Total–Variation–Diminishing Implicit–Explicit Runge–Kutta Methods for the Simulation of Double-Diffusive Convection in Astrophysics

Friedrich Kupka [a,*,1]
Natalie Happenhofer [a,2]
Inmaculada Higueras [b,3]
Othmar Koch [c,1]

[a] *University of Vienna, Faculty of Mathematics, Nordbergstraße 15, A-1090 Wien, Austria*

[b] *Universidad Pública de Navarra, Departamento de Ingeniería Matemática e Informática, Campus de Arrosadia, 31006 Pamplona, Spain*

[c] *Vienna University of Technology, Institute for Analysis and Scientific Computing, A-1040 Wien, Austria. Formerly at University of Vienna, Faculty of Mathematics, Nordbergstraße 15, A-1090 Wien, Austria*

Abstract

We put forward the use of *total–variation–diminishing* (or more generally, *strong stability preserving*) implicit–explicit Runge–Kutta methods for the time integration of the equations of motion associated with the semiconvection problem in the simulation of stellar convection. The fully compressible Navier–Stokes equation, augmented by continuity and total energy equations, and an equation of state describing the relation between the thermodynamic quantities, is semi-discretized in space by *essentially non-oscillatory schemes* and dissipative finite difference methods. It is subsequently integrated in time by Runge–Kutta methods which are constructed such as to preserve the total variation diminishing (or strong stability) property satisfied by the spatial discretization coupled with the forward Euler method. We analyse the stability, accuracy and dissipativity of the time integrators and demonstrate that the most successful methods yield a substantial gain in computational efficiency as compared to classical explicit Runge–Kutta methods.

Key words: hydrodynamics, stellar convection and pulsation, double-diffusive convection, numerical methods, total–variation–diminishing, strong stability preserving, TVD, SSP
1991 MSC: 65M06, 65M08, 65M20, 65L05, 76F65
PACS: 97.10.Cv, 97.10.Sj, 47.11.-j, 02.70.-c

Introduction

Numerical hydrodynamical simulations are a common tool in astrophysical research. Just as some of their counterparts in the atmospheric sciences and in oceanography, astrophysical fluid flows are characterized by a vast range of timescales which are present in the solutions of the dynamical equations governing the temporal evolution of such flows [23]. Large relative changes of the solutions typically occur on the hydrodynamical timescale $\tau_{fluid} = \Delta x/|\mathbf{u}|$. Here, $\mathbf{u}$ is the local flow velocity and Δx is the local spatial resolution of the simulation, which coincides with the grid size obtained from spatial discretization of the governing partial differential equations. However, some of the physically important processes can also operate on much shorter timescales than τ_{fluid}. Examples include radiative transfer, sound waves, magnetohydrodynamic processes, and chemical or nuclear reactions (see [23] for example, and references therein).

In stellar astrophysics the two most important among those timescales are that of radiative energy exchange at the scale of a grid cell, τ_{rad}, and the time τ_{sound} a sound wave needs to cross such a cell.

As long as sound waves are energetically or dynamically unimportant, a numerical simulation can be advanced with much larger time steps by using semi-implicit time integration methods, for example, by a fractional step approach (see [11] for a general introduction).

Similarly, if radiative transfer has the numerical characteristics of a stiff problem, as is the case, for instance, for numerical simulations of the surface layers of A-type stars [29], implicit time integration appears desirable as well. Another important example where radiative diffusion can limit the time-step is the numerical simulation of double-diffusive processes in stellar interiors.

Semiconvection is the most important special case of double-diffusive convection in astrophysics. Models of stellar structure and evolution predict settings where the heavier product of nuclear fusion provides stability to a zone

* Corresponding Author
Email addresses: Friedrich.Kupka@univie.ac.at (Friedrich Kupka), natalie.happenhofer@univie.ac.at (Natalie Happenhofer), higueras@unavarra.es (Inmaculada Higueras), othmar@othmar-koch.org (Othmar Koch).
URLs: http://www.mpa-garching.mpg.de/~fk/ (Friedrich Kupka), http://www.othmar-koch.org (Othmar Koch).
[1] Supported by the Austrian Science Fund (FWF), project P21742-N16
[2] Supported by the Austrian Science Fund (FWF), project P20973
[3] Supported by the Ministerio de Ciencia e Innovación, project MTM2008-00785

which otherwise would be unstable to convective overturning, because temperature sufficiently rapidly decreases against the direction of gravity. Such a zone would become convective if its composition were mixed. The question whether such a zone should be treated as if it were mixed or not is referred to as the *semiconvection problem* (see [26], [38], and Chap. 13.3 and 13-A in [47], for example). A thorough physical analysis of the semiconvection problem based on numerical simulations in two spatial dimensions for a parameter set relevant to stellar astrophysics is given in [49]. Further discussions and reviews on this topic can be found, for instance, in [4,5,24,43,46]. For this problem, long total integration times are required even when the microphysics is idealised, whereas the time integration step is governed by τ_{rad} and τ_{sound}.

For reasons outlined above, in this paper we discuss the advantages of implicit–explicit (IMEX) Runge–Kutta methods for simulations of stellar convection and diffusion in the parameter regime commonly associated with semiconvection as discussed in [49]. These methods treat only part of the right-hand sides implicitly, where the resulting (generally nonlinear) equations can be solved by means of a generalized Poisson problem. It turns out that the *total–variation–diminishing (TVD)* property is essential for a numerical time integrator to be successful in simulations of the problems in our focus: to suppress spurious oscillations in the spatial discretization (which in this paper we realize for the hyperbolic terms by essentially non-oscillatory schemes and by dissipative centered finite difference schemes for the parabolic terms), this property has been demonstrated to be necessary for a stable integration in [13].

The TVD property is more generally referred to as *strong stability preserving (SSP)* or *monotonicity* when norms other than the total variation norm or even sublinear functionals are considered. When the space discretization has the property that the functional of the discrete spatial profile is decreased in the course of numerical time propagation by the forward Euler method for a time-step $\Delta t_{\mathrm{F}E}$, then an SSP method preserves this property under a step-size restriction of the form $\Delta t \leq C\Delta t_{\mathrm{F}E}$ with $C > 0$. Since the term *total–variation–diminishing* is more commonly used in the context of astrophysical simulations, where the total–variation–seminorm is the functional of interest, we mostly use these terms here synonymously. We expect the SSP (or TVD) IMEX methods to be useful also for other astrophysical problems where a high radiative (conductive) diffusivity of internal energy (temperature) restricts the time-step of hydrodynamical simulations, such as simulations of stellar surface convection with steep temperature gradients or at high resolution.

The outline of the paper is as follows. First, we introduce SSP IMEX Runge–Kutta schemes and survey the related literature in Section 1.

Next, in Section 2 we specify the general set of equations to be solved in numerical simulations of semiconvection and related flows and describe the general

solution techniques implemented for this class of problems in the ANTARES code [35] which we use for the numerical examples discussed further below. Subsequently, we discuss how this framework has to be modified when solving the dynamical equations with the IMEX approach.

In Section 3 we analyse several SSP IMEX methods from the literature with respect to their radius of absolute monotonicity, stability and dissipativity. We show that the methods yield a significant advantage over classical explicit Runge–Kutta schemes with respect to both efficiency and accuracy and we also suggest a modification for one of the methods, which turns out to improve its efficiency.

We then present numerical simulations of semiconvection in Section 4 to demonstrate the efficiency of the SSP IMEX methods as compared to the classical, explicit SSP Runge–Kutta time integrators and some non-SSP IMEX methods from the literature by giving numerical examples for a single layer in a physical scenario similar to that one studied in [49]. We conclude this paper by a summary of the main properties of the IMEX methods, suggesting reasons for preferring particular methods and providing an outlook on interesting applications, which appear especially suited for this numerical approach.

1 Implicit–Explicit Runge–Kutta Methods for Semiconvection

To introduce our numerical methods in an abstract setting, we consider the ODE initial value problem

$$\dot{y}(t) = F(y(t)) + G(y(t)), \qquad y(0) = y_0, \tag{1}$$

where we assume that the vector fields F and G have different stiffness properties. For this type of problems, *partitioned Runge–Kutta schemes* [15], also called *additive Runge–Kutta schemes*, are popular. Methods of this kind use different Runge–Kutta formulae for the treatment of the two vector fields. We will see that the spatial semi-discretization of (15) below and the associated boundary conditions give rise to this kind of system.

An s-stage partitioned Runge–Kutta method characterized by coefficient matrices $A = (a_{i,j})$ and $\tilde{A} = (\tilde{a}_{i,j})$ defines one step $y_{\text{old}} \to y_{\text{new}}$ by

$$y_i = y_{\text{old}} + \Delta t \sum_{j=1}^{s} a_{i,j} F(y_j) + \Delta t \sum_{j=1}^{s} \tilde{a}_{i,j} G(y_j), \quad i = 1, \ldots, s, \tag{2}$$

$$y_{\text{new}} = y_{\text{old}} + \Delta t \sum_{j=1}^{s} b_j F(y_j) + \Delta t \sum_{j=1}^{s} \tilde{b}_j G(y_j). \tag{3}$$

If $a_{i,j} = 0$ for $j \geq i$, the method is referred to as an *implicit–explicit (IMEX)* method.

These methods have first been investigated with respect to the SSP or TVD property in the context of hyperbolic systems with relaxation, where $G = \frac{1}{\varepsilon}\hat{G},\ \varepsilon \ll 1$, in [36]. Ibidem, the common specification for strong stability preserving IMEX methods is introduced. An IMEX method is referred to as 'SSP$k(s, \sigma, p)$' when it has the following properties: k is the order of the method in the stiff limit $\varepsilon \to 0$, which is characterized by the coefficients for the explicit part. The latter must necessarily be SSP and is referred to as the *asymptotically SSP scheme*. s is the number of stages in the implicit scheme and σ the number of stages in the explicit scheme. p is the global order of the resulting combined method. It is essential to observe that if the implicit scheme characterized by $\tilde{A} = (\tilde{a}_{i,j})$ is a *diagonally implicit Runge–Kutta (DIRK)* method, then the explicit part is evaluated only once in each stage, providing the desired computational advantage [36].

The analysis in [36] is valid only for $\varepsilon \ll 1$ [21]. However, several useful examples of strong stability preserving IMEX Runge–Kutta methods are given, see Section 3.

[21] develops a comprehensive theory of strong stability preserving additive Runge–Kutta schemes which extends the concepts for standard Runge–Kutta methods in a natural way:

Let $\tau,\ \tilde{\tau}$ be the step-size restrictions for monotonicity of the explicit Euler method for the vector fields F and G, respectively. We define the *region of absolute monotonicity*

$$\mathcal{R}(A, \tilde{A}) = \{(r, \tilde{r}) \in \mathbb{R}^2 : (A, \tilde{A}) \text{ is absolutely monotonous on } [-r, 0]\times[-\tilde{r}, 0]\}, \tag{4}$$

where the absolute monotonicity at a point $(r_0, \tilde{r}_0)$ is characterized by algebraic relations for the matrices $A,\ \tilde{A}$. The boundary in the first quadrant, $\partial\mathcal{R}(A, \tilde{A}) \cap \{(r, \tilde{r}) : r, \tilde{r} \geq 0\}$, is denoted as the *curve of absolute monotonicity*. The significance of the region $\mathcal{R}(A, \tilde{A})$ is expressed in the following theorem [21]:

Theorem 1.1 *Let $(A, \tilde{A})$ be absolutely monotonous at $(-r, -\tilde{r})$ with step-size coefficients $\tau,\ \tilde{\tau}$. Then for $h \leq \min\{r\tau, \tilde{r}\tilde{\tau}\}$, diminishing of the norm holds,*

$$\|y_i\| \leq \|y_{\text{old}}\|, \quad i = 1, \ldots, s, \qquad \|y_{\text{new}}\| \leq \|y_{\text{old}}\|.$$

[22] gives order barriers for strong stability preserving additive Runge–Kutta methods similarly to [28]. The order of an additive Runge–Kutta method $(A, \tilde{A})$ is bounded by the orders of A and $\tilde{A}$, respectively. This implies for

IMEX methods the order barrier $p \leq 4$ [28]. Moreover, [22] gives a simple algebraic criterion for a nontrivial region of absolute monotonicity in terms of incidence matrices of A, $\tilde{A}$. Some examples of strong stability preserving IMEX Runge–Kutta methods analysed in Section 3 can be seen in [22,28].

Some other relevant issues can be studied for IMEX methods. In this paper we focus on stability regions, error constants and dissipativity analysis. We close this section with a brief description of these concepts.

The stability region of IMEX Runge–Kutta methods is defined in [1,2] via the test equation of the form (1), where

$$F(u) = \mathrm{i}\beta u, \quad G(u) = \alpha u, \qquad \alpha \leq 0 < \beta. \tag{5}$$

For this problem,

$$y_{\text{new}} = R(z) y_{\text{old}}, \qquad z = \alpha\Delta t + \mathrm{i}\beta\Delta t, \tag{6}$$

and the stability region is the part of the complex plane where $|R(z)| < 1$. We will perform the corresponding analysis of the stability function for the methods considered in Section 3.

To determine error constants of the methods we have computed the empirical convergence orders by solving the non-linear test problem

$$y'(t) = (1 + \sin(y(t))) + (y^2(t) - \sin(y(t))), \qquad y(0) = 0, \tag{7}$$

with the known exact solution $y(t) = \tan(t)$. In this paper, the error constants are determined from the errors at $t = 1.3$. Their size is vital for the comparison of the accuracy of the methods of the same order and therefore the assessment of the work/precision relation. Of course, this single example only gives a rough indication of the size of the error constant and no rigorous estimate, but it seems sufficient for our purpose of comparing the methods in our focus with respect to accuracy, which we do in Sections 4 and 5 further below. The example was chosen such as to represent a nonlinear initial value problem with known solution whose profile is arbitrarily unsmooth as $t \to \frac{\pi}{2}$.

Finally, we study the dissipativity of the time integrators in conjunction with suitable space discretizations. [45] gives a justification for considering only the diffusion term in this context since the advection term becomes negligible asymptotically. We will thus investigate the dissipativity of the implicit scheme specified by $\tilde{A}$. To this end, we apply the spatial discretizations $L_{\Delta x}$ in our focus to the heat equation

$$u_t = b u_{xx},$$

and associate for the spatial discretization $u_{j\pm k} \leftrightarrow \mathrm{e}^{\pm \mathrm{i}\theta k}$. Thus, we compute the *amplification factor* $g(\theta, \mu) = R((L_{\Delta x} u)_j)$, with $\mu = b\frac{\Delta t}{(\Delta x)^2}$. This represents

the factor by which oscillations of frequency θ are amplified in each time step. We pay particular attention to the case $\theta = \pi$ corresponding to the mesh width. If $g = 0$ or $|g| = 1$ also for a smaller $0 < |\theta_0| < \pi$, then this value would represent the limit for a robust integration. However, such a pathological behaviour is only conceivable for methods of higher order.

The spatial discretizations which were found to show a dissipative behaviour in [27] are the second-order three-point scheme (the upper index refers to the time step)

$$u_{xx}(x_j, t_n) \approx \frac{u^n_{j+1} - 2u^n_j + u^n_{j-1}}{(\Delta x)^2} =: (L_{\Delta x} u^n)_j, \tag{8}$$

and the fourth-order stencil

$$(L_{\Delta x} u^n)_j := \frac{-u^n_{j+2} + 16\, u^n_{j+1} - 30\, u^n_j + 16\, u^n_{j-1} - u^n_{j-2}}{12\,(\Delta x)^2}. \tag{9}$$

These are the methods actually implemented in ANTARES, where (9) is the default (see also [27]).

2 Solving the Hydrodynamical Equations with ANTARES

In our model, the fundamental equation of motion is the *fully compressible Navier–Stokes equation* which describes momentum conservation:

$$(\rho \mathbf{u})' + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + pI) = \rho \mathbf{g} + \nabla \cdot \sigma. \tag{10}$$

The state variables in the model equations generally depend on the spatial variables (x, y, z) and time t. In the simulations presented in Section 4 we solve problems in two spatial variables, whence the variable z will be dropped in the rest of the paper. The (explicit) dependencies are stated in Table 1. For simplicity, we omit the dependencies in the problem specification (10)–(14). The model is completed by the *continuity equation*

$$\rho' + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{11}$$

which ensures conservation of mass, and the *total energy equation*

$$e' + \nabla \cdot (\mathbf{u}(e + p)) = \rho(\mathbf{g} \cdot \mathbf{u}) + \nabla \cdot (\mathbf{u} \cdot \sigma) + Q_{\mathrm{rad}}, \tag{12}$$

which describes conservation of the latter. In the case of a two-component fluid, the system is augmented by the concentration equation of the second species,

$$(c\rho)' + \nabla \cdot (c\rho \mathbf{u}) = \nabla \cdot (\rho \kappa_c \nabla c). \tag{13}$$

The variables and parameters which appear in the model formulation are collected in Table 1.

$\rho = \rho(x, y, z, t)$	gas density
$c = c(x, y, z, t)$	concentration of second species
$\mathbf{u} = \mathbf{u}(x, y, z, t) = (u, v, w)^T$	flow velocity
$\rho\mathbf{u}$	momentum density
$\mathbf{u} \otimes \mathbf{u}$	Kronecker product
$p = p(T, \rho)$	gas pressure
$\mathbf{g} = (g, 0, 0)^T$	gravitational acceleration
$\sigma = \sigma(x, y, z, t)$	viscous stress tensor for zero bulk viscosity
η	dynamic viscosity (appears in the definition of σ)
$e = e(x, y, z, t) = e_{\text{int}} + e_{\text{kin}}$	total energy density
$T = T(x, y, z, t)$	temperature
$Q_{\text{rad}} = Q_{\text{rad}}(x, y, z, t)$	radiative source term
$c_{\text{p}} = c_{\text{p}}(T, \rho, c)$	specific heat at constant pressure
$\chi_\nu = \chi_\nu(T, \rho, c)$	(specific) opacity at frequency ν
$K = K(T, \rho, c)$	radiative (or thermal) conductivity
$\kappa_T = K/(c_{\text{p}}\rho)$	radiative (or thermal) diffusivity
$\kappa_c = \kappa_c(T, \rho, c)$	diffusion coefficient for species c
$I_\nu = I_\nu(\mathbf{r}),\ \mathbf{r} = \mathbf{r}(x, y, z)$	specific intensity along the ray of direction $\mathbf{r}$
$S_\nu = S_\nu(x, y, z)$	source function

Table 1
Variables and parameters in the equations (10)–(14).

In general, the radiative source term Q_{rad} is determined as the stationary limit of the *radiative transfer equation*

$$\mathbf{r} \cdot \nabla I_\nu = \rho\chi_\nu(S_\nu - I_\nu), \tag{14}$$

which is solved for all ray directions $\mathbf{r}$ and for all frequencies ν, resulting in the specific intensity $I_\nu(\mathbf{r})$, for details see [47]. S_ν here denotes the *source function*.

The equations of hydrodynamics (10), (11) and (12) are closed by the equation of state which describes the relation between the thermodynamic quantities. For the particular choice, see [35].

For the initial condition, a slightly perturbed static model stellar atmosphere or stellar envelope is used which is equipped with a small seed velocity field

or density perturbation to start dynamics away from equilibrium.

In the framework discussed below, boundary conditions are based on the assumption that all quantities are periodic in both horizontal directions. Moreover, for the hydrodynamical equations, "closed" (Dirichlet) boundary conditions at the upper and lower boundary of the computational domain are used. A recent development is to replace these by "open" (Robin) boundary conditions. These allow inflow and outflow of fluid along the vertical direction which is defined by the direction of $\mathbf{g}$. For the radiative transfer equation (14), incoming radiation at the boundary of the computational domain must be specified. Since double-diffusive convection in stars takes place in regions which are optically thick, the quantity Q_{rad} can accurately be obtained by means of the diffusion approximation for radiative transfer, $Q_{\mathrm{rad}} = \nabla \cdot F_{\mathrm{rad}} = \nabla \cdot (K \nabla T)$. In this case, further knowledge about the intensity I_ν is not necessary.

The ANTARES code [35] solves this system of equations numerically in either one, two, or three spatial dimensions on a rectangular grid (spherical coordinates with a logarithmically rectangular grid are also possible, i.e., the grid may be locally rectangular with logarithmic grading in the radial component).

For the spatial discretization, ANTARES allows the definition of several grids which can be nested inside each other to improve resolution in regions of interest. At the moment, ANTARES provides up to three levels of nested grids. For the hyperbolic terms, discretizations of ENO (*essentially non-oscillatory* [40]) type are implemented. These comprise classical ENO methods, WENO (*weighted essentially non-oscillatory*) methods [40] (optionally in conjunction with Marquina flux splitting [8]) and CNO (*convex non-oscillatory*) schemes [31]. Each of the methods uses adaptive stencils which are chosen such as to avoid spurious oscillations in the computed solution. The spatial derivatives are calculated for each direction separately.

The parabolic terms are discretized by dissipative finite difference schemes [27] of fourth order. The *radiative heating rate* is determined by the *short characteristics method*, or by means of the diffusion approximation for radiative transfer, where appropriate. For the time integration, *total variation diminishing* Runge–Kutta methods [39,41] are employed.

ANTARES implements two different parallelization concepts. For architectures with distributed memory, domain decomposition is used and realized by an MPI implementation. In this approach each grid is split along the horizontal direction(s) and optionally, also along vertical ones, into subdomains. The memory required to store the computational variables for each subdomain is provided by the resources available to the CPU core performing the computations necessary for that subdomain. In this way, each CPU core is mapped to a specific geometrical volume. However, since some supercomputers offer only

a limited amount of memory per CPU core and because the domain decomposition approach is not very efficient on small grids, ANTARES offers a second type of parallelization which can be used along with or independently of the former. It is based on a shared memory concept for each subdomain and is implemented through OpenMP directives. Therefore, the most time consuming operations which can also be performed independently of each other are identified and computed in parallel. This approach scales only to a moderate number of CPU cores, but allows improvement of the scaling and the computational speed of the domain decomposition based parallelization for a larger number of problems and for a greater variety of computer architectures.

In the following, the dynamical evolution of the fluid is described by the multispecies Navier–Stokes equations presented above. Additionally, dimensionless quantities such as the Prandtl number $\mathrm{Pr} = c_\mathrm{p}\eta/K$, the Lewis number $\mathrm{Le} = c_\mathrm{p}\rho\kappa_c/K$, the Rayleigh number Ra and the stability parameter R_ρ are defined to determine the diffusivities κ_T, κ_c and the viscosity η. The former quantities arise in the definition of the starting model but do not appear in the evolution equations (15) below. Since we solve the dynamical equations for a compressible flow, we specify the vertical extent of the simulation domain in multiples of the pressure scale height $H_p = P/(\rho g)$. For the simulations presented in Section 4 the domain always covers $1H_p$. In the physical model, intermolecular forces are neglected, so the fluid is assumed to be an ideal gas. The radiative source term Q_rad is modelled using the diffusion approximation with a heat conductivity K which is constant in time and otherwise only a function of the vertical coordinate [33,34]. We use a variant of this setting here, where not only Pr, Le, and c_p, but also K, κ_c, and η/ρ take constant values [49]. This setup simplifies studies of the basic physics while it is still useful for extrapolations to astrophysically relevant cases (cf. [49]).

For the model problem, the multispecies Navier–Stokes equations can be recast as

$$\underbrace{\frac{d}{dt}\begin{pmatrix}\rho\\ \rho c\\ \rho\mathbf{u}\\ e\end{pmatrix}}_{\dot{y}(t)} = \underbrace{-\nabla\cdot\begin{pmatrix}\rho\mathbf{u}\\ \rho c\mathbf{u}\\ \rho\mathbf{u}\otimes\mathbf{u}+P-\sigma\\ e\mathbf{u}+P\mathbf{u}-\mathbf{u}\cdot\sigma\end{pmatrix}-\begin{pmatrix}0\\ 0\\ \rho g\\ \rho g\mathbf{u}\end{pmatrix}}_{F(y(t))} + \underbrace{\nabla\cdot\begin{pmatrix}0\\ \rho\kappa_c\nabla c\\ 0\\ K\nabla T\end{pmatrix}}_{G(y(t))}. \quad (15)$$

In the context of the problem (15), the i^th implicit stage of an IMEX Runge–Kutta method is typically of the form

$$y_i = y^* + \Delta t\, \tilde{a}_{i,i}\, G(y_i), \tag{16}$$

where y^* is known from previous stages. This is a consequence of $\tilde{a}_{i,j} = 0$ for $j > i$.

This translates to

$$\rho_i = \rho^*, \tag{17}$$
$$(\rho c)_i = (\rho c)^* + \Delta t\, \tilde{a}_{i,i} \nabla \cdot (\rho_i \kappa_c \nabla c_i), \tag{18}$$
$$(\rho \mathbf{u})_i = (\rho \mathbf{u})^*, \tag{19}$$
$$e_i = e^* + \Delta t\, \tilde{a}_{i,i} \nabla \cdot (K \nabla T_i). \tag{20}$$

Rearranging (18) leads to

$$\frac{\rho^*}{\Delta t\, \tilde{a}_{i,i}} c_i - \nabla \cdot (\rho^* \kappa_c \nabla c_i) = \frac{\rho^* c^*}{\Delta t\, \tilde{a}_{i,i}}. \tag{21}$$

Obviously, this is a general elliptic equation for c_i of the form

$$g(x,y)\varphi(x,y) - \nabla \cdot (h(x,y) \nabla \varphi(x,y)) = f(x,y). \tag{22}$$

Due to model assumptions, (20) can also be transformed to resemble a general elliptic equation. We start by recalling

$$e = e_{\text{int}} + e_{\text{kin}} \tag{23}$$
$$= e_{\text{int}} + \frac{1}{2}\rho |\mathbf{u}|^2. \tag{24}$$

Bearing in mind equation (19), equation (20) reads

$$e_{\text{int}\ i} = e^*_{\text{int}} + \Delta t\, \tilde{a}_{i,i} \nabla \cdot (K \nabla T_i). \tag{25}$$

The equation of state for an ideal gas [4] relates the temperature T and the internal energy e_{int} via

$$e_{\text{int}} = \frac{3}{2} \frac{T \rho R_{\text{gas}}}{m}, \tag{26}$$

[4] Note that, as is common in astrophysics, the gas constant R_{gas} is taken to be relative to the atomic mass unit such that the dimensionless mean molecular weight can be used in the equation of state instead of the molar mass.

if we assume the ratio of the specific heats at constant pressure and volume to equal 5/3. Here, m denotes the mean molecular weight of the compound.

So we have at stage i

$$e_{\text{int } i} = \frac{3}{2} \frac{T_i \rho^* R_{\text{gas}}}{m_i} \tag{27}$$

and therefore, we arrive at

$$\frac{3}{2} \frac{R_{\text{gas}} \rho^*}{m_i \, \Delta t \, \tilde{a}_{i,i}} T_i - \nabla \cdot (K \nabla T_i) = \frac{e^*_{\text{int}}}{\Delta t \, \tilde{a}_{i,i}}. \tag{28}$$

Since m_i is evaluated using the mass fraction c_i, it is necessary to solve equation (21) first.

Thus, the solution of an implicit stage translates to the solution of the generalized Poisson equations for the mass fraction c and the temperature T. In the ANTARES framework, finite elements are used for the discretization of (22). The resulting linear system is solved by the conjugate gradient method. For parallel computations, the Schur complement algorithm is applied. A detailed description is given in [14].

The above procedure applies without modification to the case of the fully compressible Navier–Stokes equation. However, for low Mach number flows a splitting approach is preferable where the terms containing pressure are treated separately in a post-processing step, i. e. after evaluation of all other terms for the computation of the velocity fields. The latter are obtained for the current step from an additional generalized Poisson equation for the pressure. This procedure is described in detail in [11]. Consequently, the explicit stage is evaluated here using a fractional step method [30] as implemented in [17].

3 Strong Stability Preserving IMEX Schemes from the Literature

In this section we study different SSP IMEX methods from the literature focusing on the topics described in Section 1. The results are summarized in Figure 16 and Table 10.

All the strong stability preserving IMEX schemes listed in the following subsections have DIRK (diagonally implicit Runge–Kutta) methods as the implicit scheme. This structure ensures that the stages can be solved successively, and the explicit part only has to be evaluated once in each stage.

[36] gives an IMEX SSP2(2,2,2) method with nontrivial region of absolute monotonicity ($\gamma = 1 - \frac{1}{\sqrt{2}}$):

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{29}$$

$$A \qquad\qquad \tilde{A}$$

The coefficients imply $\mathcal{R}(A) = 1,\ \mathcal{R}(\tilde{A}) = 1 + \sqrt{2}$, and

$$\mathcal{R}(A, \tilde{A}) = \{(r, \tilde{r}) : 0 \le r \le 1,\ 0 \le \tilde{r} \le \sqrt{2}(1 - r)\},$$

see [21].

The stability region is entirely located in the left half plane, tangent to the imaginary axis and unbounded as $\Re(z) \to -\infty$. Hence, the schemes are $A(\alpha)$–stable with $\alpha = \frac{\pi}{2}$, but not A-stable [16]. Moreover, $\lim_{\Re(z)\to-\infty} R(z) = 0$. A plot of the stability region is given in Figure 1 (left).

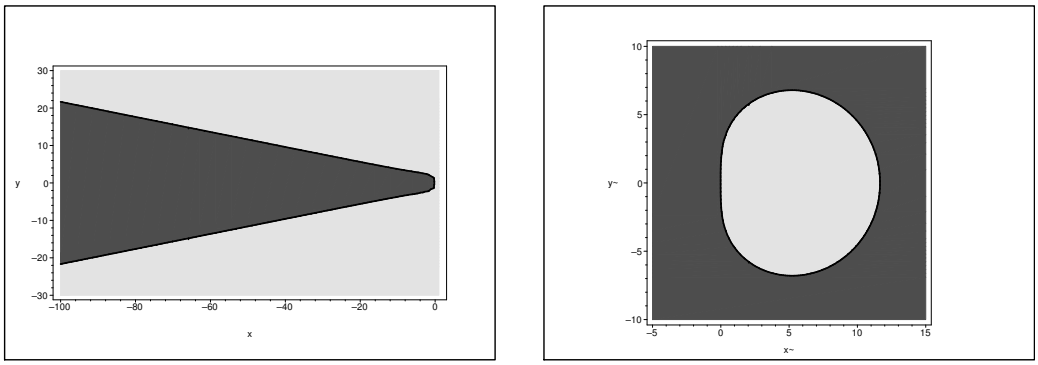


Fig. 1. Stability regions of IMEX method (29) (left) and $\tilde{A}$ (right).

The stability function of the implicit scheme $\tilde{A}$ is

$$R_{\tilde{A}}(z) = 2\,\frac{\left(1+\sqrt{2}\right)\left(1+z+\sqrt{2}\right)}{\left(2-z+\sqrt{2}\right)^2}. \tag{30}$$

A plot of the related stability region is shown in Figure 1 (right). The scheme $\tilde{A}$ appears to be A-stable and satisfies $\lim_{\Re(z)\to-\infty} R_{\tilde{A}}(z) = 0$, implying L-stability.

The dissipativity analysis for the implicit scheme defined by $\tilde{A}$ yields the amplification factors for the standard three-point space discretization (8) and the

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$2\,\frac{\left(1+\sqrt{2}\right)\left(1-2\,\mu+\mu\,\sqrt{2}+\sqrt{2}\right)}{\left(-2-2\,\mu+\mu\,\sqrt{2}-\sqrt{2}\right)^2}$
$\frac{\pi}{2}$	$2\,\frac{\left(1+\sqrt{2}\right)\left(1-2\,\mu+\sqrt{2}\right)}{\left(2+2\,\mu+\sqrt{2}\right)^2}$
π	$2\,\frac{\left(1+\sqrt{2}\right)\left(1-4\,\mu+\sqrt{2}\right)}{\left(2+4\,\mu+\sqrt{2}\right)^2}$

Table 2
Values of $g(\theta,\mu)$ for some θ, implicit scheme in (29), three point space discretization (8).

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$12\,\frac{\left(1+\sqrt{2}\right)\left(6-15\,\mu+8\,\sqrt{2}\mu+6\,\sqrt{2}\right)}{\left(-12-15\,\mu+8\,\sqrt{2}\mu-6\,\sqrt{2}\right)^2}$
$\frac{\pi}{2}$	$6\,\frac{\left(1+\sqrt{2}\right)\left(3-7\,\mu+3\,\sqrt{2}\right)}{\left(6+7\,\mu+3\,\sqrt{2}\right)^2}$
π	$6\,\frac{\left(1+\sqrt{2}\right)\left(3-16\,\mu+3\,\sqrt{2}\right)}{\left(6+16\,\mu+3\,\sqrt{2}\right)^2}$

Table 3
Values of $g(\theta,\mu)$ for some θ, implicit scheme in (29), fourth order space discretization (9).

fourth order stencil (9), respectively. The amplification factors are evaluated at the points $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \pi\}$ in Tables 2 and 3, respectively.

The first positive zero of $g(\pi,\mu)$ is ≈ 0.6035 for the three-point scheme (8), where the function changes its sign, and $|g(\pi,\mu)|$ never exceeds 1. The first positive zero of $g(\pi,\mu)$ for the fourth order space discretization (9) is ≈ 0.4526, where the function changes its sign. The modulus never exceeds 1.

Modification of γ

We may conceive of optimizing the method (29) by adapting the value of the parameter γ in the definition of $\tilde{A}$ according to the resulting stability, accuracy, and dissipativity properties. The region of absolute monotonicity depends on γ as follows [19]:

$$\mathcal{R}(\tilde{A}) = \begin{cases} \frac{1}{1-3\gamma}, & 0 \le \gamma \le \frac{1}{4}, \\ \frac{1-2\gamma}{2\gamma^2-4\gamma+1} - \sqrt{\frac{4\gamma-1}{(2\gamma^2-4\gamma+1)^2}}, & 1/4 < \gamma < 1 - \frac{1}{\sqrt{2}}, \\ 1+\sqrt{2}, & \gamma = 1 - \frac{1}{\sqrt{2}}, \\ \frac{1-2\gamma}{2\gamma^2-4\gamma+1} + \sqrt{\frac{4\gamma-1}{(2\gamma^2-4\gamma+1)^2}}, & 1 - \frac{1}{\sqrt{2}} < \gamma \le \frac{1}{2}, \end{cases} \tag{31}$$

$$\mathcal{R}(A, \tilde{A}) = \begin{cases} \left\{(r, \tilde{r}) : 0 \le r \le 1\,, 0 \le \tilde{r} \le \frac{1-r}{1-\gamma}\right\}, & 0 \le \gamma \le \frac{1}{3}, \\ \left\{(r, \tilde{r}) : 0 \le r \le \frac{1-2\gamma}{\gamma}\,, 0 \le \tilde{r} \le \frac{1-r}{1-\gamma}\right\}, & \frac{1}{3} \le \gamma \le \frac{1}{2}. \end{cases} \tag{32}$$

A plot of the function $\mathcal{R}(\tilde{A})$ in dependence of γ is given in Figure 2.

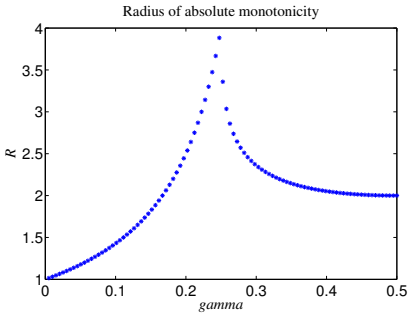


Fig. 2. Radius of absolute monotonicity $\mathcal{R}(\tilde{A})$ as a function of γ for (29).

The regions of absolute monotonicity $\mathcal{R}(A, \tilde{A})$ for the values $\gamma \in \{0.1, 0.2, 0.3\}$ are plotted in Figure 3.

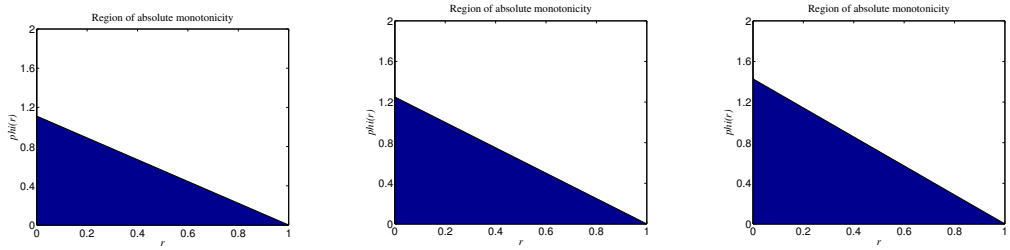


Fig. 3. Regions $\mathcal{R}(A, \tilde{A})$ for $\gamma \in \{0.1, 0.2, 0.3\}$ for (29).

The stability regions for the IMEX schemes for the different values of γ cover bounded subdomains of the left half plane for $\gamma < 0.25$, while for $\gamma \ge 0.25$, the stability regions cover unbounded domains in the left half plane. In fact, the left boundaries z_{left} satisfy

$$z_{\text{left}} = \begin{cases} \frac{2}{4\gamma-1}, & \gamma < 0.25, \\ -\infty, & \gamma \ge 0.25. \end{cases}$$

However, even in the cases with unbounded stability regions, in general $\lim_{\Re(z)\to-\infty} R(z) \neq 0$. The boundaries of the stability regions are plotted in Figure 4 (left), where values equal to 0 represent unbounded stability regions. The implicit schemes $\tilde{A}$ show the same stability behaviour concerning both the boundaries of the stability regions and the limits for $\Re(z) \to -\infty$.

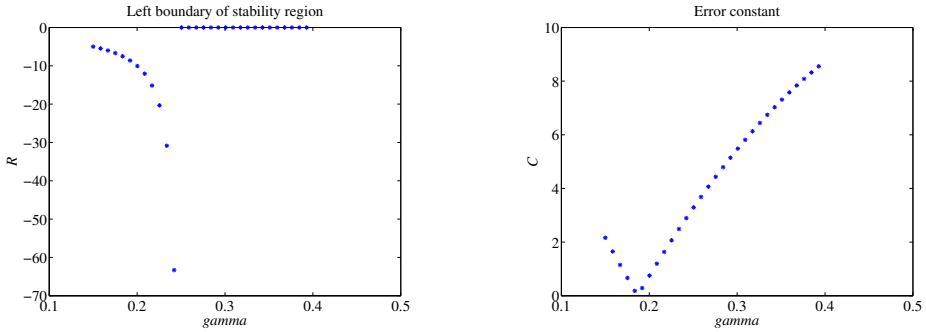


Fig. 4. Left boundary of the stability region (left) and error constant computed for (7) (right) as a function of γ for (29).

It was demonstrated with a MATLAB implementation that the convergence order two is retained also for the modified values of γ. The error constant depends on γ, however. In Figure 4 (right) we plot the error constant as a function of γ, where the error is determined at $t = 1.3$ for the test problem (7). We note that for small γ, the error constant decreases as γ grows, while for $\gamma > 0.1833$ the constant grows monotonically. This behaviour does not appear to be related to the results we obtain for the dissipativity analysis, where for $\gamma = 0.25$ the behaviour changes.

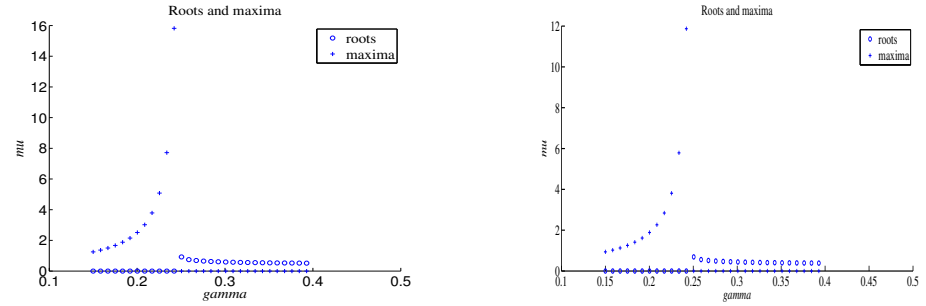


Fig. 5. Dissipativity analysis for (29) for different values of γ, three point space discretization (8) (left) and fourth order discretization (9) (right).

To assess the dissipativity properties of the modified scheme we vary γ and compute the first positive zero of $g(\pi, \mu)$ and the point μ where the modulus of this function exceeds 1.

For $\gamma \geq 0.25$, the amplification factor for the three-point space discretization (8) has a zero at $\mu = \frac{1-2\gamma-\sqrt{4\gamma-1}}{8\gamma^2-16\gamma+4}$, but there is no real root for $\gamma < 0.25$. Also for $\gamma < 0.25$, $|g|$ exceeds 1 at $\mu = \frac{1}{2-8\gamma}$, and this behaviour does not occur for

$\gamma \geq 0.25$. For the discretization (9), the behaviour is the same, but the scale with respect to μ is multiplied by 0.75 (see also Section 3.2 in [27]).

To illustrate this analysis, in Figure 5 we vary γ and compute for 30 values of γ spaced equidistantly in the interval $[0.15, 1.1 - 1/\sqrt{2}]$ the first positive zero of $g(\pi, \mu)$ and the point μ where the modulus of this function exceeds 1. This analysis is given for the three point space discretization (8) in Figure 5. We conclude that it may be of interest to choose $\gamma < 0.25$ to ensure $0 < g(\pi, \mu) < 1$. This can be achieved by a value of γ just slightly smaller than 0.25.

Taking into account the results displayed in Fig. 4 we suggest to optimize γ by taking it as large as necessary to avoid any linear stability restrictions due to the term $G(y(t))$ in (1) and as small as possible to minimize the stability constant C. Note that some special values for γ are 0, $(1-1/\sqrt{3})/2 \approx 0.2113$, and $1/4$, in addition to the originally proposed value of $1 - 1/\sqrt{2} \approx 0.2929$. They are readily identified to yield the classical explicit SSPRK(2,2) method [5] of order two by Shu & Osher [41], the optimal implicit third order SSP method with two stages [10], and the optimal implicit second order SSP method with two stages [10]. In our numerical tests reported in Section 4 below, we found that the choice $\gamma = 0.24$ yielded the most efficient time integrator.

An SSP2(3,3,2) Method

[21] gives an IMEX SSP2(3,3,2) method with nontrivial region of absolute monotonicity:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & & & \\ A & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{5} & \frac{1}{5} & 0 & 0 \\ \frac{3}{10} & \frac{1}{10} & \frac{1}{5} & 0 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & & & \\ \tilde{A} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \tag{33}$$

This is a modification of a scheme from [36], where the latter turned out to have a trivial region of absolute monotonicity. It holds that $\mathcal{R}(A) = 2$ and $R(\tilde{A}) = \frac{5}{9}(\sqrt{70} - 4)$, and

$$\mathcal{R}(A, \tilde{A}) = \{(r, \tilde{r}) : 0 \leq r \leq 1,\ 0 \leq \tilde{r} \leq \phi(r)\},$$

where

$$\phi(r) = \left\{ \frac{1}{4}(-28 + 9r) + \frac{1}{4}\sqrt{1264 - 984r + 201r^2} \right\}.$$

[5] We will henceforth use the common specification 'SSPRK(s,p)' introduced in [28] for an s-stage order p explicit strong stability preserving Runge–Kutta method.

We note that the latter is a correction with respect to [21], since we have found r to be necessarily bound by 1 in $\mathcal{R}(A, \tilde{A})$. A plot of $\mathcal{R}(A, \tilde{A})$ is given in Figure 6.

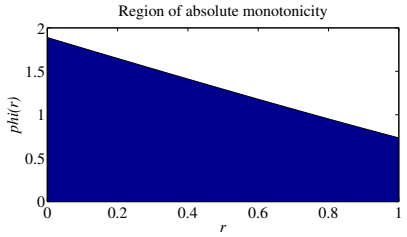


Fig. 6. Region of absolute monotonicity for method (33).

A plot of the stability region is given in Figure 7 (left). We observe that the stability region is tangent to the imaginary axis and appears to be unbounded as $\Re(z) \to -\infty$. Moreover, $\lim_{\Re(z)\to-\infty} R(z) = 0$. The stability function of the

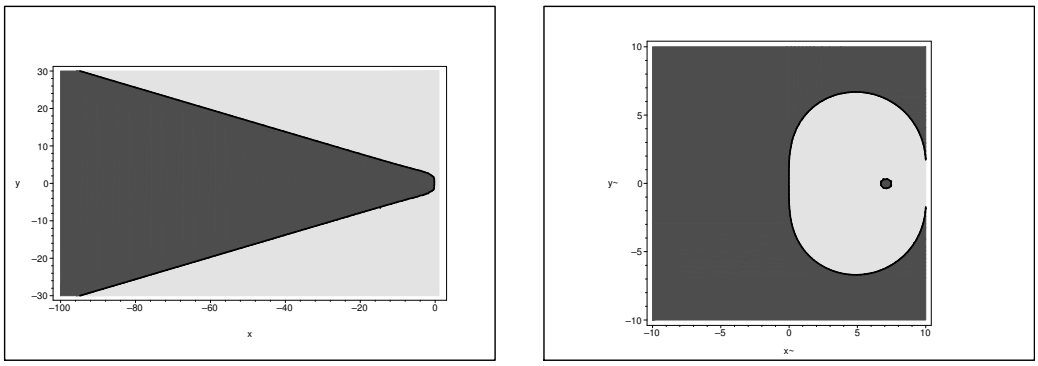


Fig. 7. Stability region of IMEX method (33) (left) and $\tilde{A}$ (right).

implicit scheme is given by

$$R_{\tilde{A}}(z) = \frac{-150 - 40\,z + 9\,z^2}{2\,(-5+z)^2\,(-3+z)}. \tag{34}$$

A plot is shown in Figure 7 (right). Note that the stability region is not connected. The method is A-stable, however. Moreover, the scheme $\tilde{A}$ satisfies $\lim_{\Re(z)\to-\infty} R_{\tilde{A}}(z) = 0$.

The dissipativity analysis for the implicit scheme defined by $\tilde{A}$ yields the amplification factors for the standard three-point space discretization (8) and the fourth order stencil (9). These amplification factors are evaluated at the points $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \pi\}$ in Tables 4 and 5, respectively.

The first positive zero of $g(\pi, \mu)$ is ≈ 0.6064 for the three-point space discretization (8), where the function changes its sign, and $|g(\pi, \mu)|$ never exceeds 1. The first positive zero of $g(\pi, \mu)$ is ≈ 0.4551 for the fourth order space discretization (9), where the function changes its sign, and $|g(\pi, \mu)|$ never exceeds 1.

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$-\frac{75+20\,\mu\,\sqrt{2}-40\,\mu-27\,\mu^2+18\,\mu^2\sqrt{2}}{\left(-5+\mu\,\sqrt{2}-2\,\mu\right)^2\left(-3+\mu\,\sqrt{2}-2\,\mu\right)}$
$\frac{\pi}{2}$	$-\frac{-75+18\,\mu^2+40\,\mu}{(5+2\,\mu)^2(3+2\,\mu)}$
π	$-\frac{-75+80\,\mu+72\,\mu^2}{(5+4\,\mu)^2(3+4\,\mu)}$

Table 4
Values of $g(\theta,\mu)$ for some θ, implicit scheme in (33), three point space discretization (8).

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$-9\,\frac{1800-1200\,\mu+640\,\mu\,\sqrt{2}-1059\,\mu^2+720\,\mu^2\sqrt{2}}{\left(-30-15\,\mu+8\,\mu\,\sqrt{2}\right)^2\left(-18-15\,\mu+8\,\mu\,\sqrt{2}\right)}$
$\frac{\pi}{2}$	$-9/2\,\frac{-450+280\,\mu+147\,\mu^2}{(7\,\mu+15)^2(7\,\mu+9)}$
π	$-9\,\frac{320\,\mu+384\,\mu^2-225}{(15+16\,\mu)^2(9+16\,\mu)}$

Table 5
Values of $g(\theta,\mu)$ for some θ, implicit scheme in (33), fourth order space discretization (9).

An SSP3(3,3,3) Method

[22] gives the following SSP3(3,3,3) method with nontrivial region of absolute monotonicity:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} \qquad A \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & \frac{14}{15} & \frac{1}{15} & 0 \\ \frac{1}{2} & \frac{7}{30} & \frac{1}{5} & \frac{1}{15} \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} \qquad \tilde{A} \tag{35}$$

It holds $\mathcal{R}(A)=1$ and $\mathcal{R}(\tilde{A})=\frac{5}{47}(13-2\sqrt{7})$, and

$$\mathcal{R}(A,\tilde{A})=\{(r,\tilde{r}):0\le r\le 1,\ 0\le\tilde{r}\le\phi(r)\},$$

where

$$\phi(r)=\frac{15}{302}\left(28-25r-\sqrt{180-192r+21r^2}\right).$$

A plot of $\mathcal{R}(A,\tilde{A})$ is given in Figure 8.

The stability region occupies a bounded domain in the negative half-plane, and slightly overlaps the imaginary axis, see Figure 9 (left). Note that

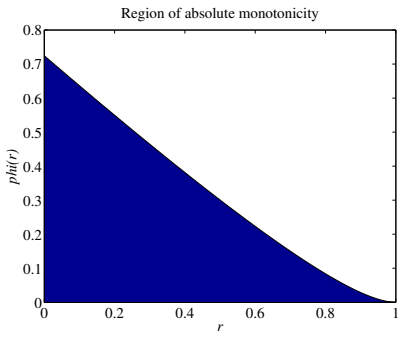


Fig. 8. Region of absolute monotonicity for method (35).

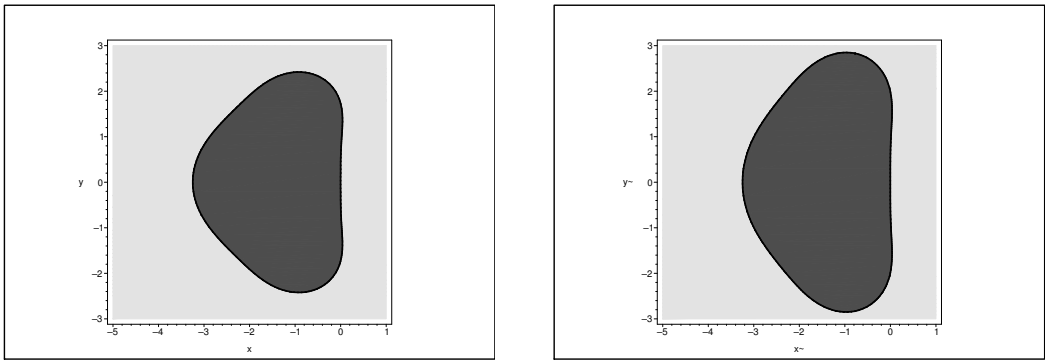


Fig. 9. Stability regions of IMEX method (35) (left) and implicit method $\tilde{A}$ (right).

$\lim_{\Re(z)\to-\infty} |R(z)| = \infty$. The stability function of the implicit scheme is given by

$$R_{\tilde{A}}(z) = \frac{450 + 390\,z + 167\,z^2 + 47\,z^3}{2\,(-15+z)^2}. \tag{36}$$

A plot is shown in Figure 9 (right). The point where the stability region intersects the negative real half-line is located at $x \approx -3.248$ for the IMEX scheme. The same value is computed for the stability region of the implicit scheme. However, the stability region of $\tilde{A}$ extends further along the imaginary axis. Finally, $\lim_{\Re(z)\to-\infty} |R_{\tilde{A}}(z)| = \infty$.

The dissipativity analysis for the implicit scheme defined by $\tilde{A}$ yields the amplification factors for the standard three-point space discretization (8) and for the fourth order stencil (9). These amplification factors are evaluated at the points $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \pi\}$ in Tables 6 and 7, respectively.

For the three-point space discretization (8), the first positive zero of $g(\pi, \mu)$ is ≈ 0.4650, where the function changes its sign, and $g(\pi, \mu) = -1$ for $\mu \approx 0.8122$. The first positive zero of $g(\pi, \mu)$ is ≈ 0.3488 for the fourth-order space discretization (9), where the function changes its sign, and $g(\pi, \mu) = -1$ at $\mu = 0.6093$.

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$\frac{225+195\,\mu\,\sqrt{2}-390\,\mu+501\,\mu^2-334\,\mu^2\sqrt{2}+329\,\mu^3\sqrt{2}-470\,\mu^3}{\left(-15+\mu\,\sqrt{2}-2\,\mu\right)^2}$
$\frac{\pi}{2}$	$-\frac{-225+390\,\mu-334\,\mu^2+188\,\mu^3}{\left(15+2\,\mu\right)^2}$
π	$-\frac{1504\,\mu^3+780\,\mu-1336\,\mu^2-225}{\left(15+4\,\mu\right)^2}$

Table 6
Values of $g(\theta,\mu)$ for some θ, implicit scheme in (35), three point space discretization.

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$1/12\,\frac{97200-210600\,\mu+112320\,\mu\,\sqrt{2}+353706\,\mu^2-240480\,\mu^2\sqrt{2}-429345\,\mu^3+301928\,\mu^3\sqrt{2}}{\left(-90-15\,\mu+8\,\mu\,\sqrt{2}\right)^2}$
$\frac{\pi}{2}$	$-1/6\,\frac{-12150+24570\,\mu-24549\,\mu^2+16121\,\mu^3}{\left(45+7\,\mu\right)^2}$
π	$-1/3\,\frac{-6075+28080\,\mu-64128\,\mu^2+96256\,\mu^3}{\left(45+16\,\mu\right)^2}$

Table 7
Values of $g(\theta,\mu)$ for some θ, implicit scheme in (35), fourth order space discretization.

4 Numerical Experiments

In this section, we present the results of the simulations performed with the time integrators discussed in this paper and compare their performance to the classical explicit 2-stages second order and 3-stages third order methods of Shu & Osher [41], and the 3-stages second order method of [28]. They are the optimum SSP RK methods for their given number of stages and order and we refer to them by their common technical specifications 'SSPRK(2,2)' and 'SSPRK(3,3)', as well as 'SSPRK(3,2)', respectively. The coefficients of the SSPRK(2,2) and SSPRK(3,3) methods were derived for a different purpose in [18] and in [9], where the third order method was proposed as an embedding formula for the second order method (see also [3]). Shu & Osher [41] derived a first framework for deducing higher order Runge–Kutta methods with the total variation diminishing property and first identified the optimal explicit second and third order methods with two and three stages, respectively. For these SSPRK(2,2) and SSPRK(3,3) schemes an analysis of their stability, dissipativity and accuracy properties can be found in [27]. For the SSPRK(3,2) method, a similar study is given in the Appendix section further below. Note that SSPRK(2,2), SSPRK(3,2), and SSPRK(3,3) are the explicit schemes in the IMEX methods (29), (33), and (35), respectively.

Furthermore, we also consider some non-SSP methods from the literature,

both explicit and IMEX, for the sake of comparison, and also an asymptotically stable SSP IMEX scheme from [36].

Implementation Issues

To define the test problem, according to [33] we specify a hydrostatic configuration which is unstable against convection. The simulation of a single semiconvective layer requires the mean molecular weight to be linearly (and stably) stratified. As time evolves, we expect convection to set in and mix the zone completely, although its development is inhibited by the stable mean molecular weight gradient. A critical quantity in this process is hence the buoyancy timescale and we return to this topic further below.

The simulations shown here have been performed on the Vienna Scientific Cluster, using 64 CPU cores in parallel. The spatial resolution is 400×400 grid points. Simulation time is measured in units of *sound crossing times (scrt).* One scrt is defined as the time taking an acoustic wave to propagate from the bottom to the top of the simulated box. In our simulations, $1\,\mathrm{scrt} = 5215.5\,s$ and the simulation time is 200 scrt.

Restrictions on the time-step Δt are imposed by heat diffusion τ_T, diffusion of the second species τ_c, the viscosity τ_{visc} and the velocity of the fluid τ_{fluid},

$$\Delta t = \min\{\tau_c, \tau_T, \tau_{\mathrm{visc}}, \tau_{\mathrm{fluid}}\}, \tag{37}$$

where

$$\tau_c = \frac{C_c}{\kappa_c}\,\min\{(\Delta x)^2, (\Delta y)^2\}, \quad \tau_T = \frac{C_T}{\kappa_T}\,\min\{(\Delta x)^2, (\Delta y)^2\},$$
$$\tau_{\mathrm{visc}} = \frac{C_{\mathrm{visc}}\,\rho}{\eta}\,\min\{(\Delta x)^2, (\Delta y)^2\}, \quad \tau_{\mathrm{fluid}} = \frac{C_{\mathrm{fluid}}}{\max(|\mathbf{u}|)}\,\min\{\Delta x, \Delta y\},$$

with (not necessarily equal) Courant numbers (CFL numbers) C_c, C_T, C_{visc}, C_{fluid}. This assumes that the time-step limitation due to sound waves has been removed by a fractional step approach as mentioned at the end of Section 2. Otherwise, τ_{fluid} additionally depends on the sound speed. We note that the source term in $F(y(t))$ in (15), which represents buoyancy forces acting on the flow, can be neglected in the limit where $\max\{\Delta x, \Delta y\} \to 0$, since its contributions are of lower order (see [45] for a discussion of the treatment of lower order terms in stability analyses).

Due to (15), IMEX methods treat the terms $\nabla\cdot(\rho\kappa_c\nabla c)$ and $\nabla\cdot(K\nabla T)$ implicitly, so the restrictions τ_c and τ_T do not have to hold. Since at least the first part of simulations of semiconvection is usually dominated by diffusion processes and the Prandtl and Lewis numbers satisfy $\mathrm{Pr} < 1$, $\mathrm{Le} < 1$ in a

stellar context, the simulations are initially restricted by τ_T. Hence, IMEX methods can provide the desired computational advantage.

To enhance the stability and the efficiency of our methods we have implemented a heuristic to adaptively select the time steps. The most effective criterion for regulating the time steps turned out to be to monitor two-point instabilities appearing in the conservative variables $(\rho, \rho c, \rho\vec{u}, Et)$. Due to the gravitational force operating vertically, such instabilities are prone to appear in the horizontal direction.

To detect the occurrence of such oscillations, for each variable we use the difference between two grid cells to determine the sign of the corresponding gradient. In case of the density ρ this reads

$$
\begin{aligned}
d_1 &= \rho_{i,j-1} - \rho_{i,j-2}, \\
d_2 &= \rho_{i,j} - \rho_{i,j-1}, \\
d_3 &= \rho_{i,j+1} - \rho_{i,j}, \\
d_4 &= \rho_{i,j+2} - \rho_{i,j+1},
\end{aligned}
$$

where $1 < i < n_x$ and $1 < j < n_y$, assuming the grid consists of $n_x \times n_y$ points. If the sign pattern of (d_1, d_2, d_3, d_4) corresponds to $(+, -, +, \pm)$, $(-, +, -, \pm)$, $(\pm, +, -, +,)$ or $(\pm, -, +, -)$, we have located a two-point instability. Since the time-step is initially chosen to be that one required for a fully explicit time integration method as in (37), such patterns are smoothed out rapidly, if present in the initial condition. Consequently, their later occurrence is a good indicator for an instability developing because of too large a time-step taken during the time integration.

The time-step control permits the occurrence of $n_y \cdot 0.1$ two-point instabilities for fixed i. If this limit is exceeded, the time-step is repeated using a step-size decreased by a factor $\frac{2}{3}$.

To permit the system to readjust after reducing the time-step no modifications of Δt are made for the next 15 time-steps regardless of the number of two-point instabilities. If the number of oscillations still exceeds the given limit after those 15 time steps, the time-step is again reduced. If no or very few two-point oscillations are encountered for over 50 successive time-steps, the time-step is augmented by a factor of $\frac{5}{4}$.

Alternatively to this heuristic control, we could also monitor the rate of change in the solution to adjust the time-steps. However, the proper rate required to prevent the development of two-point instabilities for the present application turned out to be too pessimistic to achieve CFL numbers as high as discussed below. Furthermore, a simple control of the residual did not yield a satisfactory

Method	$\Delta t_{\max}$	Δt_{mean}	$\text{CFL}_{\max}$	CFL_{mean}	$\text{CFL}_{\text{start}}$
Singlelayer Pr = 0.1, Le = 0.1, $R_\rho = 1.1$, $\text{Ra}^* = 160000$					
SSPRK(2,2)	3.71 s	3.71 s	0.2	0.2	0.2
SSPRK(3,2)	9.31 s	9.31 s	0.5	0.5	0.5
IMEX SSP2(2,2,2)	22.21 s	11.56 s	1.20	0.62	0.2
IMEX SSP2(2,2,2), γ=0.24	36.35 s	19.44 s	1.96	1.05	0.2
IMEX SSP2(2,2,2), γ=0.24	36.35 s	19.43 s	1.96	1.05	0.3
IMEX SSP2(2,2,2), γ=0.24	37.02 s	19.45 s	2.00	1.05	0.4
IMEX SSP2(2,2,2), γ=0.24	35.53 s	19.47 s	1.91	1.05	0.5
IMEX SSP2(3,3,2)	74.52 s	57.37 s	4.02	3.09	0.4
IMEX SSP2(3,3,2)	93.15 s	57.16 s	5.02	3.08	0.5
SSPRK(3,3)	3.71 s	3.71 s	0.2	0.2	0.2
IMEX SSP3(3,3,3)	15.14 s	10.14 s	0.82	0.55	0.2
Singlelayer Pr = 0.5, Le = 0.1, $R_\rho = 1.1$, $\text{Ra}^* = 160000$					
SSPRK(2,2)	3.72 s	3.72 s	0.2	0.2	0.2
SSPRK(3,2)	9.31 s	9.31 s	0.5	0.5	0.5
IMEX SSP2(2,2,2)	23.14 s	13.84 s	1.24	0.74	0.2
IMEX SSP2(2,2,2), γ=0.24	23.13 s	15.72 s	1.24	0.85	0.2
IMEX SSP2(2,2,2), γ=0.24	22.76 s	15.79 s	1.22	0.85	0.3
IMEX SSP2(2,2,2), γ=0.24	20.32 s	15.81 s	1.09	0.85	0.4
IMEX SSP2(2,2,2), γ=0.24	22.81 s	15.85 s	1.23	0.84	0.5
IMEX SSP2(3,3,2)	40.65 s	33.54 s	2.19	1.80	0.4
IMEX SSP2(3,3,2)	40.65 s	33.87 s	2.19	1.82	0.5
SSPRK(3,3)	3.72 s	3.72 s	0.2	0.2	0.2
IMEX SSP3(3,3,3)	15.05 s	9.70 s	0.81	0.52	0.2

Table 8
Comparisons of time steps and CFL-numbers over the first 80 scrt for the case where Pr = 0.1 and over the entire 200 scrt for the case where Pr = 0.5 (see also text).

behaviour. This leaves the heuristic time-step control as the most effective method for the IMEX based time integration of our simulations.

Numerical Results with SSP Schemes

Tables 8 and 9 sum up the performance achieved with the presented Runge–Kutta schemes. The evolution of the time steps is illustrated graphically in Figures 10 and 11 for the two simulation scenarios we have focused on. Since the capability of the method is best judged in that part of the simulation where the fluid velocity is too small to severely limit the time-step Δt, the CFL-numbers and time-steps listed in Table 8 have been measured in this regime. For the first test scenario this corresponds only to the first 80 scrt (note the slope beginning after about 100 scrt in Figure 10).

In Table 8 we compare the largest possible time steps $\Delta t_{\max}$, the average time steps Δt_{mean}, the maximal and average CFL numbers resulting from the adaptive step selection, and the initial CFL numbers. The tests were performed for Prandtl numbers $\text{Pr} = 0.1$ and $\text{Pr} = 0.5$ distinguishing Simulations 1 and 2, respectively, a Lewis number $\text{Le} = 0.1$, $R_\rho = 1.1$, and a modified Rayleigh number $\text{Ra}^* = 160000$ related to the Rayleigh number through $\text{Ra}^* = \text{Ra} \cdot \text{Pr}$.

Comparing the performance of the second order schemes it is obvious that the modification of γ in IMEX SSP2(2,2,2) has a stunning effect on the stability of the scheme. The positive definiteness of dissipation in this method proves most effective in suppressing oscillations, permitting an average time-step and CFL-number up to two thirds higher than the original IMEX SSP2(2,2,2) method.

Table 8 shows that IMEX SSP2(3,3,2) permits a time-step and CFL number more than twice as high as IMEX SSP2(2,2,2) even with modified γ. However, a comparison of the computation time given in Table 9 shows that this does not improve by the same factor as the CFL-number, since as the time-step Δt grows, the iterative solver for the generalized elliptic problems requires more iterations to converge, resulting in an increase in computation time. A comparison of the computation times shows that, although IMEX SSP2(3,3,2) permits impressively large time steps, the need to solve three additional generalized Poisson problems related to the third stage takes its toll, whence the method's performance is inferior to IMEX SSP2(2,2,2) with modified γ and in case of Simulation 2, is not even competitive to SSPRK(3,2), though it still performs better than the SSPRK(2,2) scheme. Note that for $\text{Pr} = 0.1$, the IMEX methods outperform even the best explicit method, while for the more moderate $\text{Pr} = 0.5$, the best explicit integrator SSPRK(3,2) is slightly more efficient whereas the classical methods noticeably lag behind.

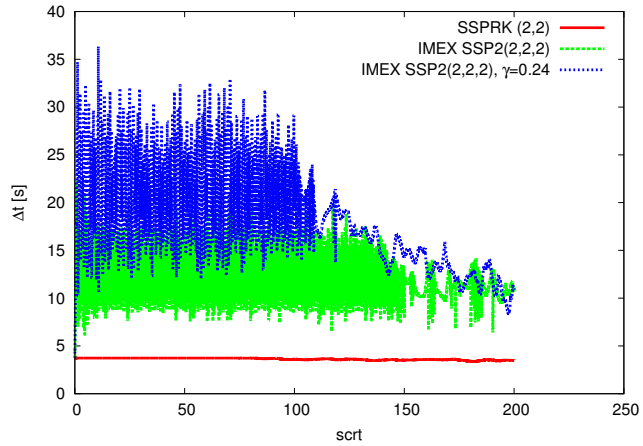
Interestingly, the initial (preset) CFL number has a negligible influence on the actual CFL number reached in the diffusive part of the simulation. However, as soon as the fluid velocity seriously restricts Δt, a higher initial CFL number leads to a significantly larger average time-step and reduces the required

Method	CFL_{start}	computation time	number of time steps
Singlelayer Pr = 0.1, Le = 0.1, $R_\rho = 1.1$, $Ra^* = 160000$			
SSPRK(2,2)	0.2	7:17:32	287 788
SSPRK(3,2)	0.5	4:40:24	115 691
IMEX SSP2(2,2,2)	0.2	6:04:54	92 492
IMEX SSP2(2,2,2), γ=0.24	0.2	4:41:15	63 586
IMEX SSP2(2,2,2), γ=0.24	0.3	4:18:59	56 741
IMEX SSP2(2,2,2), γ=0.24	0.4	4:10:26	54 015
IMEX SSP2(2,2,2), γ=0.24	0.5	4:12:38	53 941
IMEX SSP2(3,3,2)	0.4	4:31:12	27 673
IMEX SSP2(3,3,2)	0.5	4:27:46	24 266
SSPRK(3,3)	0.2	10:55:40	288 607
IMEX SSP3(3,3,3)	0.2	8:19:39	104 789
Singlelayer Pr = 0.5, Le = 0.1, $R_\rho = 1.1$, $Ra^* = 160000$			
SSPRK(2,2)	0.2	7:01:44	286 011
SSPRK(3,2)	0.5	4:34:35	114 409
IMEX SSP2(2,2,2)	0.2	5:10:01	75 384
IMEX SSP2(2,2,2), γ=0.24	0.2	4:45:36	66 997
IMEX SSP2(2,2,2), γ=0.24	0.3	4:42:00	68 591
IMEX SSP2(2,2,2), γ=0.24	0.4	4:42:24	66 847
IMEX SSP2(2,2,2), γ=0.24	0.5	4:50:52	66 828
IMEX SSP2(3,3,2)	0.4	4:43:27	31 225
IMEX SSP2(3,3,2)	0.5	4:43:26	31 112
SSPRK(3,3)	0.2	10:45:43	286 006
IMEX SSP3(3,3,3)	0.2	7:27:01	108 852

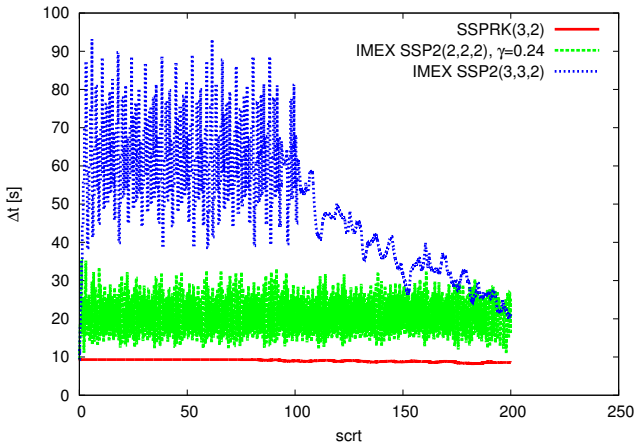
Table 9
Comparisons: computation times and overall number of time steps over 200 scrt.

computation time, since its value is used to define the time-step restriction for the terms integrated with the explicit part of the IMEX scheme.

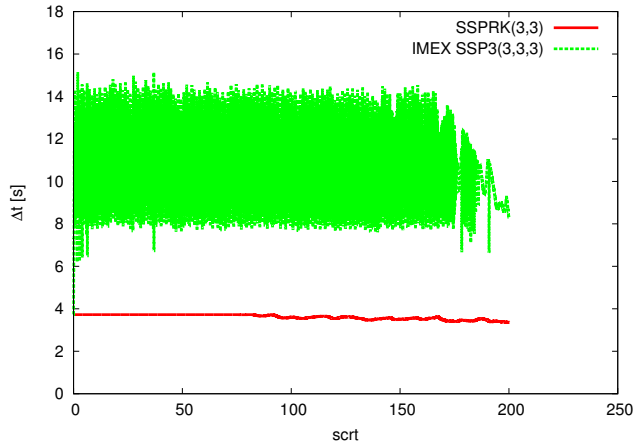
Once the time-step is limited by τ_{fluid}, the remaining gain in Δt by the IMEX schemes in ANTARES is essentially due to the optimization of the time-steps by the algorithm explained further above. Since Δt is rather small in that case



(a) $\mathrm{CFL}_{\mathrm{start}} = 0.2$

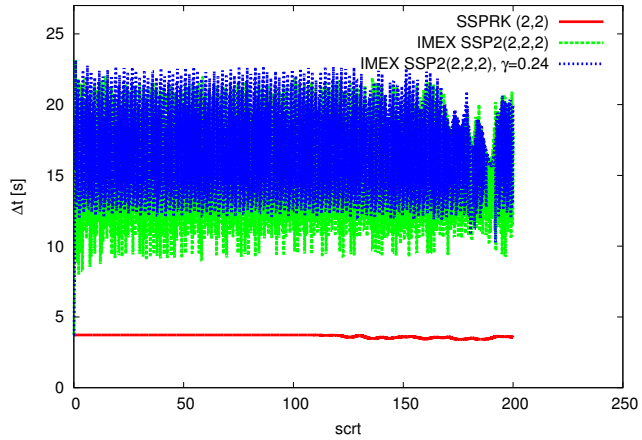


(b) $\mathrm{CFL}_{\mathrm{start}} = 0.5$

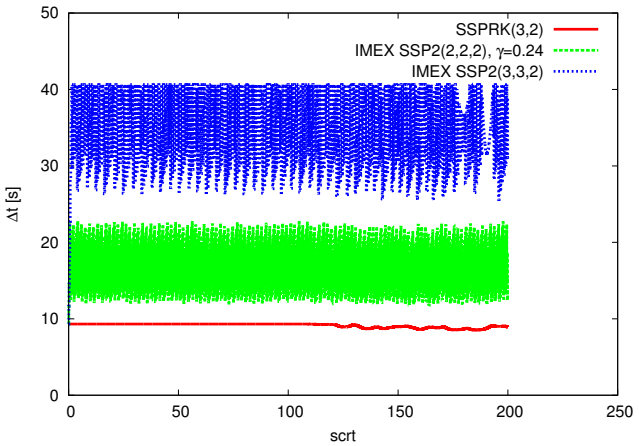


(c) $\mathrm{CFL}_{\mathrm{start}} = 0.2$

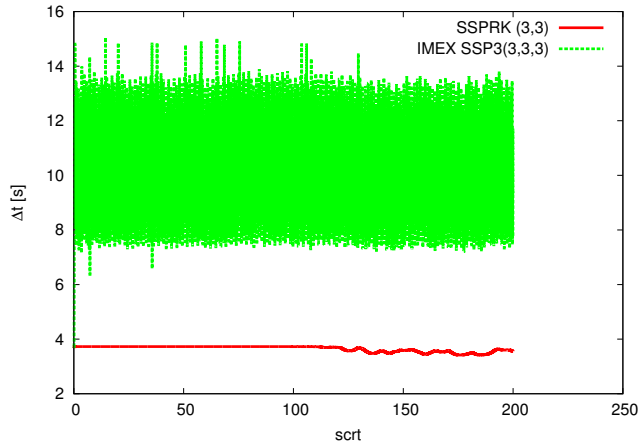
Fig. 10. Time-step evolution over 200 scrt in Simulation 1 (see text for definitions). Pictures (a) and (b) compare the time-step Δt of the second order schemes whereas (c) shows the evolution of Δt using SSPRK(3,3) and IMEX SSP3(3,3,3).



(a) $\mathrm{CFL}_{\mathrm{start}} = 0.2$



(b) $\mathrm{CFL}_{\mathrm{start}} = 0.5$



(c) $\mathrm{CFL}_{\mathrm{start}} = 0.2$

Fig. 11. Time-step evolution over 200 scrt in Simulation 2 (see text for definitions). Pictures (a) and (b) compare the time-step Δt of the second order schemes whereas (c) shows the evolution of Δt using SSPRK(3,3) and IMEX SSP3(3,3,3).

the convergence of the generalized Poisson solver is fast enough to allow the IMEX schemes to lag only slightly behind their explicit counterparts.

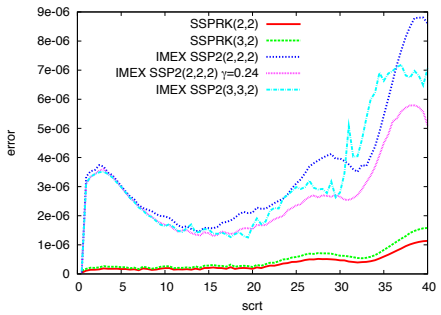
We point out that optimization of the solver for the generalized Poisson equation (21) has the potential for a further significant decrease of computation time required, both in the diffusive regime, but also if the time-step is limited by τ_{fluid}. This is important since changing the time integration method during a simulation run is not advisable if both the diffusive and the convective phase should be interpreted in a consistent manner.

A comparison of SSPRK(3,3) and IMEX SSP3(3,3,3) also shows that the larger time-steps of the semi-implicit method lead to a significant gain in computational efficiency in case of a third order method.

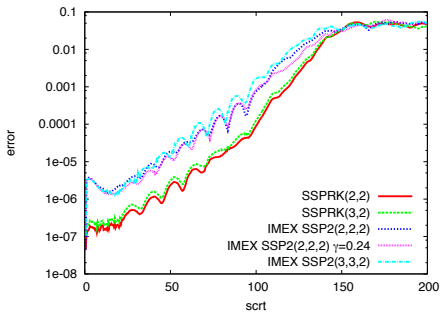
We have investigated the accuracy of the time integration with IMEX schemes by a comparison to a reference solution obtained with the SSPRK(2,2) method. The reference solution was computed on the same spatial grid but with a time-step eight times smaller than that one mentioned in Tables 8 and 9 for this method, i.e. for a CFL-number of 0.025. Figure 12 shows the root mean square difference of the mass ratio He / (H + He), obtained by summation over all grid points and normalization relative to their number, between the numerical solutions of the second order SSP IMEX methods and the reference solution for each case. For the IMEX SSP2(2,2,2) method both the results for the standard choice of $\gamma = 1 - 1/\sqrt{2}$ and the best performing value of 0.24 are displayed. We also show the normalized root mean square differences between the reference solution and the SSP second order explicit methods computed with their standard CFL number given in Table 8.

For both Simulation 1 and 2 one can easily spot the initial increase of the error due to the growing time step for the IMEX methods induced by the automatic time step control. A plateau is reached once the time-step stabilizes around a typical mean value (cf. also Figures 10 and 11). Note that the IMEX SSP2(2,2,2) method with $\gamma = 0.24$ has a smaller error than the original IMEX SSP2(2,2,2) method, as expected from the error constant shown in Figure 4. The largest differences occur for the IMEX SSP2(3,3,2) method which also has the largest mean time-step. The error constants of the different methods (see also the summarizing Table 10 below) provide a rough measure for comparing simulation runs with similar time-steps.

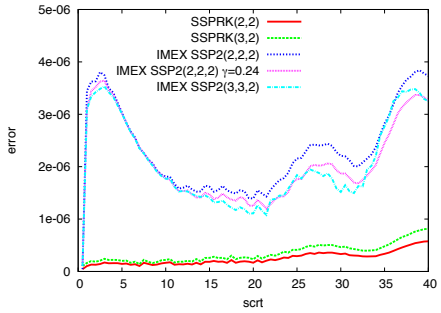
However, a comparison for a given point in time has only limited meaning. One of the reasons is that the solution changes its nature as a function of time. Initial vertical oscillations are damped out (at least first few scrt), then the velocity field slowly starts building up (visible after 15 scrt), followed by the formation of large scale gravity waves (oscillatory behaviour of the error in the range between 25 and 100 scrt) until the waves start to break



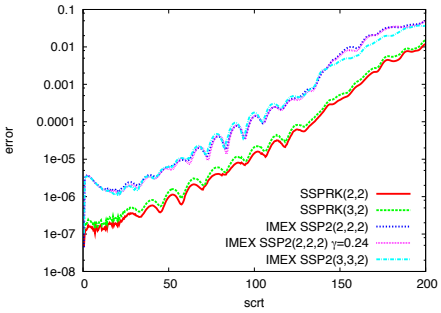
(a) Simulation 1, first 40 scrt



(b) Simulation 1, entire run



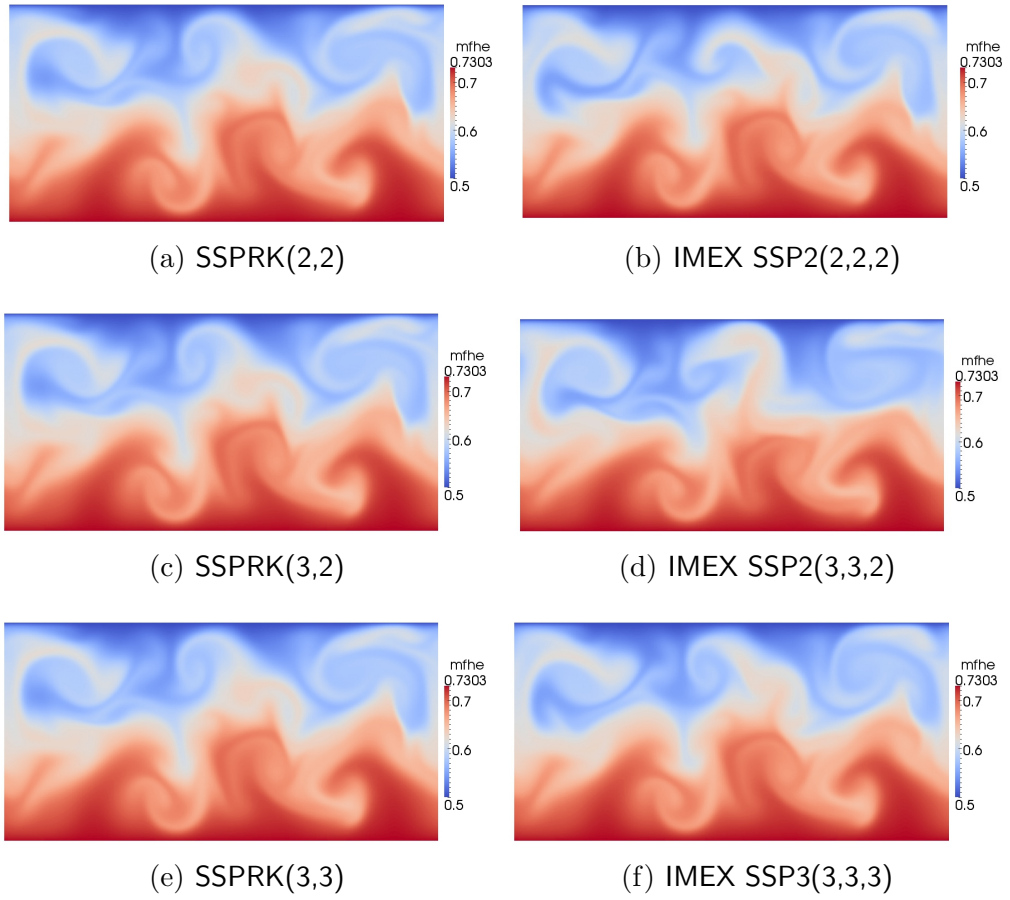
(c) Simulation 2, first 40 scrt



(d) Simulation 2, entire run

Fig. 12. Time development of the root mean square difference per grid point of the mass ratio He / (H + He) between a reference solution with the SSPRK(2,2) method and very small time-step (CFL-number 0.025) and various explicit and IMEX SSP methods of second order. In the top row, picture (a) displays the first 40 scrt on a linear scale for Simulation 1 and picture (b) shows the results for the entire run on a logarithmic scale. In the bottom row, pictures (c) and (d) show equivalent results for the case of Simulation 2.

and turbulence sets in. The importance of the contributions of each of the dynamical equations also changes during this development. The whole process leads to an increasing error until a statistically stationary, turbulent state is reached. For Simulation 1 this occurs at around 150 scrt. For Simulation 2 this is just about to occur shortly after the end of the simulation time of 200 scrt (the delay of the time development in this case is caused by the larger viscosity of the fluid which follows from the choice of Pr and Ra^*). In the turbulent state of the system spatial correlation is lost on very short timescales (a few scrt). Thus, also the reference solution no longer has a meaning due to the chaotic behaviour of the solution. The spread of the error and also the saturation value observed during this phase (picture (b) of Figure 12) is set by the Dirichlet vertical boundary conditions on the concentration c. The third order methods IMEX SSP3(3,3,3) and SSPRK(3,3) behave analogously.



(a) SSPRK(2,2) (b) IMEX SSP2(2,2,2)

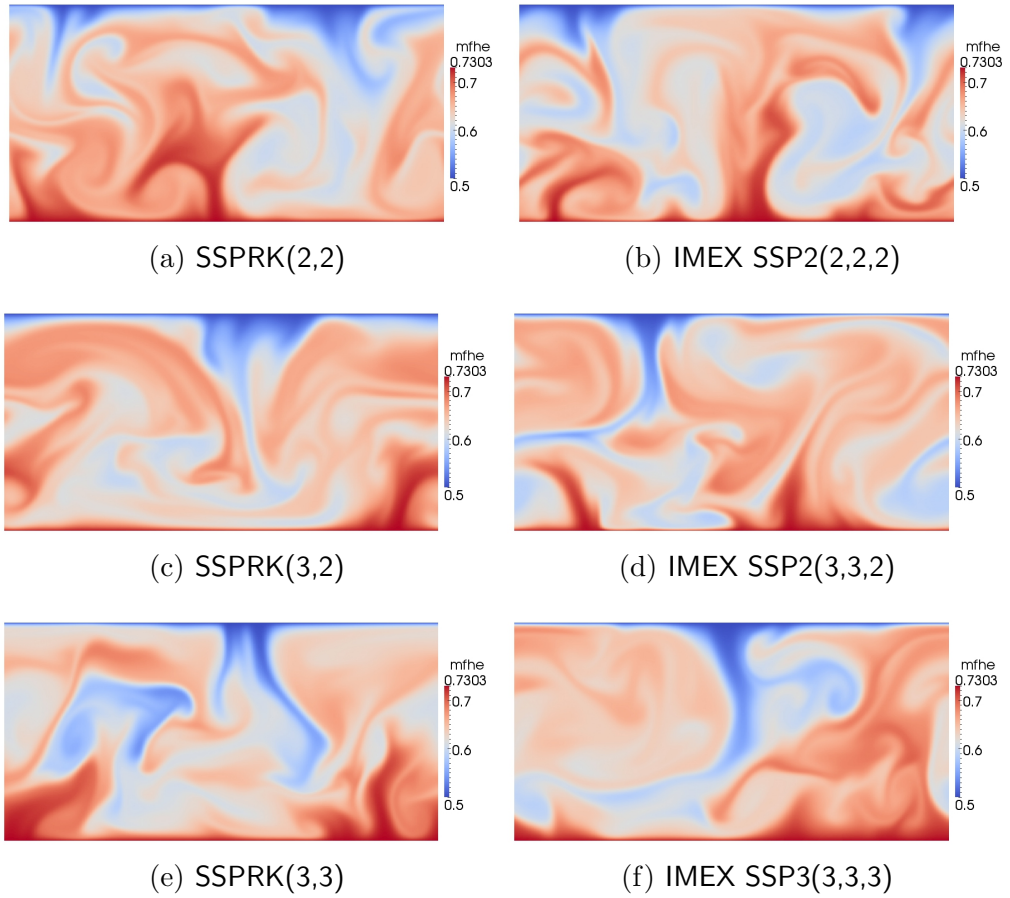
(c) SSPRK(3,2) (d) IMEX SSP2(3,3,2)

(e) SSPRK(3,3) (f) IMEX SSP3(3,3,3)

Fig. 13. Simulation 1 at $t = 125$ scrt.

To further illustrate that the large time steps of the IMEX methods during the diffusive phase do not degrade the accuracy of the time development of the solution, we compare the simulation results obtained by the different time integrators in Figures 13 and 14. The pictures show the mass ratio of He vs. He+H at each spatial point at a given instant in time. Figure 13 demonstrates that after the end of the diffusive phase, just at the onset of turbulence, which occurs at ~ 125 scrt for this problem, the results are quite comparable. The most visible differences can be found for the case of the simulation with the largest time-steps (picture (d)). It also shows that the root mean square errors displayed in Figure 12 are negligible on a qualitative (and rough quantitative) level as long as they are smaller than about 10^{-3}. As expected, however, some time after the onset of turbulence the solutions necessarily have drifted apart. This is demonstrated in Figure 14, where the solutions already look different from each other.

Recalling Figure 12 it is not surprising to find the largest differences in Figure 13 for the schemes having the largest time-steps during the diffusive phase. Still, the large scale structures of the solution begin to diverge only once the turbulent phase has been reached which in turn for each of the different time



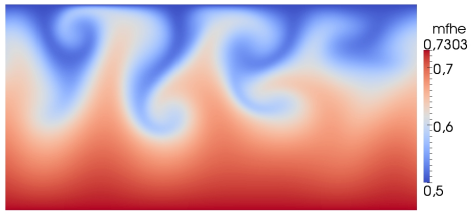
(a) SSPRK(2,2) (b) IMEX SSP2(2,2,2)

(c) SSPRK(3,2) (d) IMEX SSP2(3,3,2)

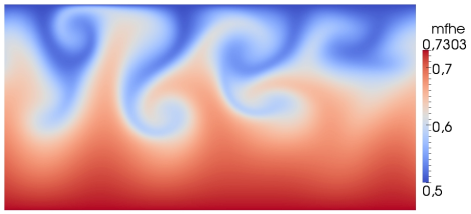
(e) SSPRK(3,3) (f) IMEX SSP3(3,3,3)

Fig. 14. Simulation results corresponding to those in Figure 13 at $t = 200$ scrt.

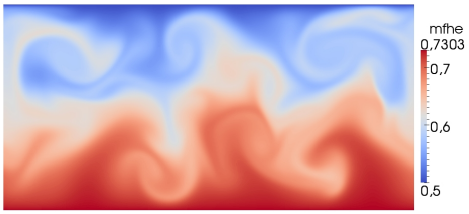
integration methods occurs after about the same integration time t. Large time-steps during the diffusive phase hence yield acceptable accuracy. Indeed, the spatial resolution is more important than the temporal one. This can be demonstrated by using a high resolution grid of 800×800 points. We have performed such a reference run for the case of Simulation 1 with the SSPRK(2,2) method for time integration. Note that the doubling of resolution leads to a four times smaller time-step during the diffusive phase. Looking at the first row (pictures (a) and (b) of Figure 15) the differences at 100 scrt, i.e. just after the onset of wave-breaking and the beginning of the turbulent phase, are still small: in the simulation with higher spatial resolution the breaking tips are somewhat more pronounced and the contrasts sharper. This has changed at 125 scrt shown in the second row of Figure 15. The common initial condition may still be inferred, but the simulations have notably evolved away from each other. Clearly, the spatial resolution is much more important than the influence of the time steps and the time integration method chosen, since picture (c) of Figure 15 is essentially indistinguishable from its counterpart with standard time resolution, picture (a) of Figure 13. Furthermore, at 100 scrt the simulation using the IMEX2(3,3,2) method for time integration, which has the largest time-step during the diffusive phase, is nearly indistinguish-



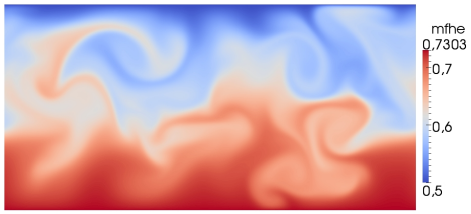
(a) high time resolution, t = 100 scrt



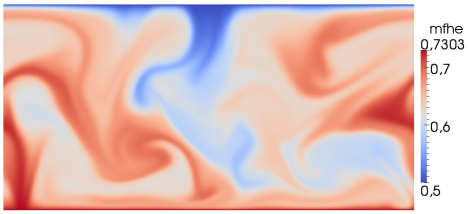
(b) high spatial resolution, t = 100 scrt



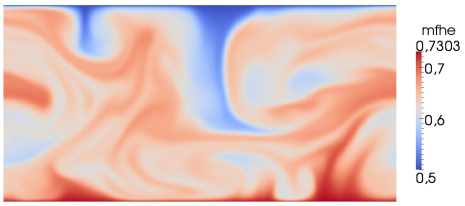
(c) high time resolution, t = 125 scrt



(d) high spatial resolution, t = 125 scrt



(e) high time resolution, t = 200 scrt



(f) high spatial resolution, t = 200 scrt

Fig. 15. Simulation 1 with SSPRK(2,2) time integration and high temporal resolution (left column) as well as high spatial (and temporal) resolution (right column). The three different rows show the results at different time t in units of scrt.

able from the SSPRK(2,2) run with high time resolution shown in picture (a) of Figure 15 (the IMEX results are not shown here for the case of 100 scrt, since the differences to the latter picture are very difficult to spot).

We conclude that the spatial resolution is indeed more important than high temporal resolution and large time steps during the diffusive phase are clearly tolerable for simulations of astrophysical convective flows, if they do not affect stability. Resolutions of 800 grid points per spatial direction in 3D are usually not affordable anyway and sometimes even large parametric studies in 2D may still be too expensive (for instance, for the case of semi-convection and purely explicit time integration methods). We note that the necessity of a resolution of 400 points, which has been used for most of the simulation runs shown here, was calculated following [48] and [49] where in turn the physical arguments of [43] had been used to estimate the thickness of solute and thermal boundary layers in semi-convection and the applicability of this approach

to the parameter range we are interested in had been confirmed. Thus, for the parameters of the more demanding one of our models, Simulation 1, we concluded the smallest structures of interest, the solutal boundary layers, to span 6 grid points, if the whole box is discretized by 400 points in each direction. Examples for them are the top and bottom boundary layers which can easily be seen in Figures 14 and 15. Indeed, in the simulation with a high spatial resolution of 800 points in each direction these layers are hardly any thinner than in the case of 400 points (cf. the bottom row of Figure 15 which shows the simulations developed well into their final, turbulent state).

The duration of the diffusive phase is determined by the stability of the stratification, parametrized through R_ρ, and related to the buoyancy term, the second term on the right-hand side of (15) and thus also the second term of $F(y(t))$ in the same equation. Timescales related to this term can be computed for a variety of physical problems. They include the reciprocal of the growth rate of small density perturbations when a fluid of higher density ρ_2 is layered above fluid of lower density ρ_1 (this situation is denoted as *Rayleigh-Taylor instability*, see [6]). In this case the growth rate follows the dispersion relation $\omega^2 = gk(\rho_2 - \rho_1)/(\rho_2 + \rho_1)$ for a local gravitational acceleration g and a perturbation with wave number k. Its magnitude can be bounded by the simple relation $\omega^2 = gk$. Similar dispersion relations are found for gravity waves and growth rates of convective instability (see the classical paper [7] and also [26] for a summary). As implied by $\omega^2 = gk$ and the form of (15), one can estimate a buoyancy time-step restriction by $t_{\text{buoy}} = \min\{(\Delta x)^{1/2}\}/g^{1/2}$ (see also [32]). For the present simulations we find $t_{\text{buoy}} \sim 360$ s, i.e. about 0.069 scrt. Note that this is about four times larger than the largest time-step reported in Table 8. Though irrelevant in the asymptotic limit and not a constraining quantity here, one might still consider this term to be integrated implicitly in other cases. However, if the excitation and breaking of waves is important to describe the onset of the turbulent convective flow, as in the present case, damping of such waves by an implicit time integration may be undesired. Hence, explicit time integration of the buoyancy term could be preferred for physical reasons, even if the time-step were actually constrained by such a splitting for the time integration.

Numerical Results with non-SSP Schemes

From Tables 8 and 9 one can readily see that during the diffusion dominated phase the SSP IMEX schemes achieve time-steps which exceed the region of absolute monotonicity ensured by (32) for IMEX SSP2(2,2,2) and their counterparts given after (33) and (35) for IMEX SSP2(3,3,2) and IMEX SSP3(3,3,3), respectively. One might hence question whether the property of absolute monotonicity is really necessary for the time integration of the numerical simulations we have considered above. To show that this property is

indeed required we have performed several test runs for the case of Simulation 1 with time integration schemes which do not diminish the total–variation norm.

The first candidate we have investigated is the ARK3(2)4L[2]SA scheme proposed in [25]. This is an L-stable, stiffly accurate third order, additive Runge–Kutta method with four stages. With its choice of coefficients it belongs to the group of IMEX methods. However, neither its explicit nor its implicit part are strong–stability–preserving (the Butcher arrays feature negative coefficients, cf. Theorem 4.2 in [28]), hence also the entire method does not fulfill the criteria of Theorem 1.1 on absolute monotonicity. If the SSP–property were of no importance, this method should be quite robust. However, it turns out that this is not the case when we use it to integrate Simulation 1 in time. Taking $\mathrm{CFL}_{\mathrm{start}}$ to be 0.2 as for the other IMEX methods presented in Table 9, the time integration with ARK3(2)4L[2]SA crashes after just 78 time-steps. Indeed, $\mathrm{CFL}_{\mathrm{start}}$ has to be lowered to 0.1 to successfully launch the simulation. But over the first 10 scrt $\mathrm{CFL}_{\mathrm{max}}$ is found to never exceed ~ 0.2 and the average $\mathrm{CFL}_{\mathrm{mean}}$ is only ~ 0.15. In conclusion the size of the time-steps achieved with this method have been found to not exceed those achieved with the explicit, three-stage, third-order SSPRK(3,3) method of [41]. Compared to IMEX SSP2(3,3,2), the maximum and mean CFL numbers are 25 and 20 times smaller, respectively. We have thus given up this simulation run after 10 scrt: evidently, the ARK3(2)4L[2]SA method is not efficient for the kind of problems we are interested in. The implicit, L-stable and stiffly accurate nature of this scheme is not sufficient to provide any advantages on its own during the diffusive phase of the simulation.

To further investigate the importance of the strong–stability–preserving property we selected the classical, explicit, third order, three–stage Runge–Kutta method first proposed in [18] and known as Heun's third order method. This is a non-SSP scheme since not all of its stages are used in the final integration which yields y_{new}, as pointed out in [28], where it was used to illustrate the growth of solutions measured in standard norms for both parabolic and hyperbolic problems in situations where the exact solution is not growing in these norms. By comparison the SSPRK(3,3) scheme was found to not exhibit such growth. When we apply Heun's third order scheme to integrate Simulation 1, we achieve a stable simulation over the entire extent of 200 scrt with the same average time-step and with the same Courant number as for the SSPRK(3,3) scheme. However, during the diffusive phase in Simulation 1 the solution itself is slowly growing in time while a velocity field is being built up by the convective instability and at the grid scale the dissipation is provided by the parabolic terms during the entire simulation.

The semi-convection problem discussed here is a rather benign example for numerical simulations in astrophysical applications. If we apply the same scheme

to a simulation of solar surface convection as in [35], i.e. for a case of 219×159 grid points and a standard, moderately low resolution of 18.57×40 km^2, and a standard choice for the microphysics with non-grey radiative transfer for the calculation of $Q_{\rm rad}$, the differences in stability become apparent. For this physical problem the term representing viscous dissipation in the momentum equation does not provide sufficient dissipation on the grid-scale at any resolution achievable in the foreseeable future and the dissipation properties of the temporal and spatial discretization of the advection operator become important (the term implicit large eddy simulation is used in such cases). While the SSPRK(3,3) method and also the SSPRK(2,2) method, used together with the spatial discretization of [40], have no problems in completing a simulation of 20 scrt with a CFL number of 0.25, Heun's third order method leads to a crash after 9.2 scrt. [6] Such failures are usually caused by numerically induced fluctuations in the low density and temperature region of the simulation box which result in negative or at least unphysically small values, whence they fall outside the tabulated region of microphysical properties. We consider this finding as sufficient to exclude non-SSP methods from being recommendable for numerical simulations of stellar convection, since also here we have chosen a rather benign test case within its class. Numerical simulations of stellar surface convection in white dwarfs, A-type stars, or Cepheids reach far more extreme conditions with up to three times higher, super-sonic Mach numbers, density contrasts around shock fronts higher by an order of magnitude and more, and for the case of A-stars and Cepheids, at lower effective resolution because of four to ten times steeper gradients and limited computational resources.

We finish our study of non-SSP Runge–Kutta schemes with a third case: an IMEX Runge–Kutta method where both the explicit and the implicit part are strong–stability–preserving, but the combined scheme does not fulfill the criteria for absolute monotonicity in the sense of Theorem 1.1. The scheme was proposed in [36] (Table IV, page 139) and indeed the IMEX SSP2(3,3,2) scheme (33) is a modification of that scheme proposed in [21] to obtain a nontrivial region of absolute monotonicity. The original scheme of [36] differs from that one only by having the entries $\{1/4, 0, 0\}$ and $\{0, 1/4, 0\}$ in the first two rows of the Butcher array of the implicit scheme instead of $\{1/5, 0, 0\}$ and $\{1/10, 1/5, 0\}$, respectively. This scheme performs quite well. Indeed, during the diffusion dominated phase of Simulation 1 its time-steps are even 6% to 8% larger than those achieved with IMEX SSP2(3,3,2), while during the turbulent phase they fall back to at most the size achieved by the scheme (33). Once more, however, we recall that astrophysical simulations often have to deal with limited resolution at least in part of the simulation domain. To reproduce such a case we have run Simulation 1 with both (33) and the original

[6] We would like to thank H. Grimm–Strele for performing the 2D solar convection simulations with ANTARES to test the stability properties of Heun's non-SSP explicit third order Runge–Kutta scheme.

scheme of [36] for the case of only 100×100 grid points while leaving everything else unchanged. In this case the boundary layer due to concentration c is represented vertically only by one to two grid points (see the discussion on resolution and reference solutions further above). While during the diffusive phase no major differences become apparent, the behaviour found for the advection–dominated, turbulent phase was discrepant: whereas nothing suspicious occurred for the scheme (33), the time-step of the scheme of [36] dropped to arbitrarily small values after ~ 113 scrt, which indicates the occurrence of two-point instabilities, and the simulation had to be terminated.

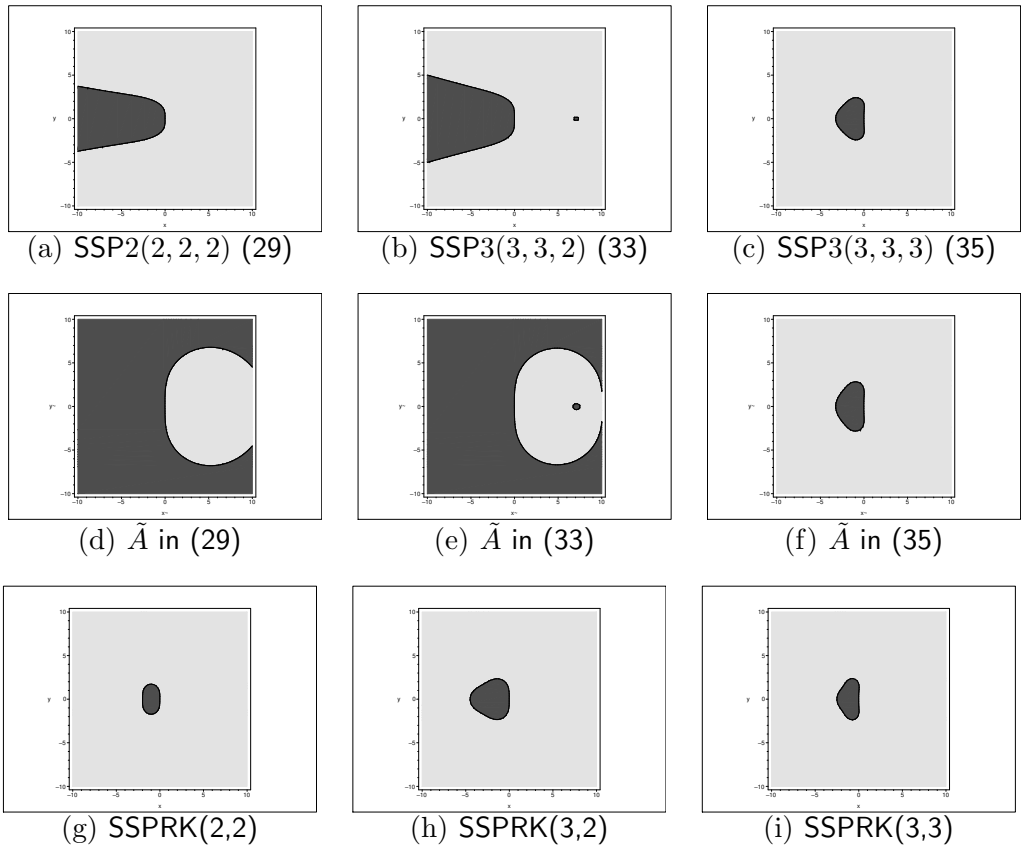
We conclude that only Runge–Kutta methods which are strong–stability–preserving have the necessary prerequisites for stable time integrations of astrophysical convection simulations. If, in addition, the time-step is limited by diffusion processes, this limitation can be overcome by IMEX methods provided their explicit and implicit parts are strong–stability–preserving. To ensure stability also in cases of low resolution IMEX SSP methods should also have a nontrivial region of absolute monotonicity as required by Theorem 1.1.

5 Conclusions and Outlook

In this paper we have given an extensive discussion of the mathematical properties and practical usefulness of total–variation–diminishing implicit–explicit Runge–Kutta methods for the time integration of advection–diffusion equations arising in the simulation of double–diffusive convection in astrophysics. In this section, we summarize the results obtained in Sections 3 and 4 (stability, dissipativity, accuracy and efficiency), and give a brief outlook on future developments.

The stability regions for the IMEX methods, their implicit sub-parts and the explicit schemes are given for comparison in Figure 16. The left boundaries of the stability regions z_{left}, the points where the amplification factors from the dissipativity analysis become zero and their moduli exceed 1, and the error constants C for all the methods we have investigated are summed up in Table 10.

We found that methods introduced in [20,21,22,36] excel over the classical explicit methods [41]. It was found that among explicit schemes, only the explicit SSPRK(3,2) scheme first proposed in [28] is competitive in situations where explicit time integration can be expected to yield sufficient efficiency and accuracy. Examples for such a scenario include simulations of solar granulation at moderately high spatial resolution, where the time-step limitation associated with diffusion is negligible (see [35] for results on this problem obtained with the ANTARES code and [44] for a review on the underlying physics).



(a) SSP2$(2,2,2)$ (29) (b) SSP3$(3,3,2)$ (33) (c) SSP3$(3,3,3)$ (35)

(d) $\tilde{A}$ in (29) (e) $\tilde{A}$ in (33) (f) $\tilde{A}$ in (35)

(g) SSPRK(2,2) (h) SSPRK(3,2) (i) SSPRK(3,3)

Fig. 16. Stability regions for IMEX methods (first row) and their implicit and explicit sub-parts (second and third row).

From Figure 16 and Table 10, clearly there is no single scheme which features the most advantageous properties in all considered aspects. However, we found in numerical experiments that the most efficient method seems to be (29) with the choice $\gamma = 0.24$. This value deviates from the optimal value for strong stability, but leads to a scheme with favourable dissipativity, stability, and accuracy properties. Depending on the domain of stability required for a given problem the value of γ in (29) may be optimized such that it is sufficiently large for stability, but small enough to minimize the error constant, while showing favourable dissipation properties (a strictly positive amplification factor with modulus less than 1 for any wave number k other than zero).

We note that for numerical problems arising from a method of lines approach to the equations of hydrodynamics, as discussed in this paper, lower order methods usually have sufficient efficiency to be competitive, since the spatial discretization limits the overall accuracy. Hence, the best explicit scheme we have tested for this kind of application is SSPRK(3,2), as it permits the largest CFL numbers among methods of this class at an affordable computational cost and with sufficient accuracy. By comparison, the classical methods of second and third order [41] offer the convenience of being usable together as an

embedding formula. However, this approach is more than twice as expensive as SSPRK(3,2), as can be seen from a comparison with SSPRK(3,3) using Table 9.

Method	z_{left}	$g_{4\text{th}} = 0$	$\lvert g_{4\text{th}}\rvert = 1$	C
IMEX SSP2(2,2,2)	$-\infty$	0.452^*	—	5.17
IMEX SSP2(2,2,2), $\gamma = 0.24$	-50	—	9.375	2.79
IMEX SSP2(3,3,2)	$-\infty$	0.455^*	—	8.05
IMEX SSP3(3,3,3)	-3.248	0.348^*	0.609	11.6
Forward Euler	-2	0.187^*	0.375	12.6
SSPRK(2,2)	-2	—	0.375	16.2
SSPRK(3,2)	-4.519	0.672^*	0.847	6.40
SSPRK(3,3)	-2.512	0.299^*	0.471	22.8

Table 10
Summary of the analysis of SSP integrators. The asterisk in the third column indicates a change of sign at μ for $g_{4\text{th}}(\pm\pi, \mu)$. Other details are given in the text.

We have also demonstrated that the larger time-steps achieved by SSP IMEX methods reduce the accuracy of the solution during the diffusive phase of the semi-convection simulations by an acceptably small amount. For the applications shown here, and indeed for a majority of astrophysical fluid dynamical simulations, the accuracy is limited by spatial resolution (and thus eventually by existing computational resources) while the time-steps are limited by stability. This makes IMEX methods attractive, since quite often the most severe limitations stem from stiff terms representing diffusion processes (for restrictions due to sound waves other operator splitting based methods are existing). However, as we have shown by a comparison with results from non-SSP methods, it is important that the IMEX methods are strong–stability–preserving to maximize stable time-steps no matter whether the constraint is due to the implicitly integrated terms (diffusion) or the explicitly integrated ones (advection). From that point of view SSP IMEX methods with a non-trivial region of absolute monotonicity as defined by Theorem 1.1 are the most robust methods, because they allow achieving optimally large time-steps also at low resolution. We note here that while the region of absolute monotonicity has to be observed with respect to the explicitly integrated advection operator of the dynamical equations, stable time integration can be performed with step sizes falling outside it, if the restriction is due to the implicitly integrated diffusion terms. Thus, the class of optimal integrators for this kind of problem is probably larger than that of the SSP IMEX methods. However, none of the other time integration methods could significantly outperform them with respect to the time-steps achievable, and we have always found at least one case, where

competing methods fell substantially short or even failed.

There is potential to further optimize the implementation of IMEX methods. The additional computational effort due to the implicit subpart is compensated for by accuracy and stability, but could be reduced in the future by replacing the solver for the linear equations associated with the arising generalized Poisson problem by a multigrid solver. In the Boussinesq approximation, additional solution of a Helmholtz equation is necessary instead. This widens the choice of fast solvers for the system of linear equations introduced through implicit time integration. The benefits expected from faster solvers would allow taking full advantage of the potential of method (33) that is implied by the large time steps reported in Table 8. Such an improvement would likewise be useful for the present problem to minimize the overhead by any of the implicit schemes in the regime where the time-step is limited by τ_{fluid} rather than τ_T.

Acknowledgements

We gratefully acknowledge the help of H. Grimm–Strele who performed the 2D solar convection simulations with ANTARES to test the stability properties of Heun's non-SSP explicit third order Runge–Kutta scheme. We also thank H.J. Muthsam for useful discussions on the ANTARES code as well as J. Ballot and H. Grimm–Strele for help in a generic implementation of the SSPRK(3,2) method which can now be used by all modules of the code.

Appendix

The Explicit SSP Scheme of IMEX SSP2(3,3,2)

We also give the corresponding results for the explicit SSPRK(3,2) scheme A from (33), since in Section 4 we show it to excel in its practical value over the classical explicit SSP Runge–Kutta schemes schemes [41]. This scheme was first published in [28] and later declared the optimal second order scheme with three stages in [42] as well as in [37], and independently also in [12].

The stability function is

$$R_A(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{12}.$$

The stability region where $|R(z)| < 1$ occupies a bounded region in the negative half-plane, and is tangent to the imaginary axis, see Figure 17. Note that $\lim_{\Re(z)\to-\infty} |R(z)| = \infty$.

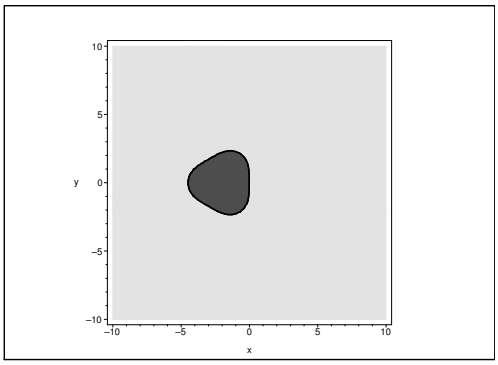


Fig. 17. Stability region of method SSPRK(3,2) (scheme A from (33)).

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$1+\mu\sqrt{2}-2\,\mu+3\,\mu^2-2\,\mu^2\sqrt{2}+7/6\,\mu^3\sqrt{2}-5/3\,\mu^3$
$\frac{\pi}{2}$	$1-2\,\mu+2\,\mu^2-2/3\,\mu^3$
π	$1-4\,\mu+8\,\mu^2-16/3\,\mu^3$

Table 11
Values of $g(\theta,\mu)$ for some θ, SSPRK(3,2) scheme (explicit scheme A from (33)), three point space discretization (8).

θ	$g(\theta,\mu)$
0	1
$\frac{\pi}{4}$	$1-5/2\,\mu+4/3\,\mu\sqrt{2}+\frac{353}{72}\,\mu^2-10/3\,\mu^2\sqrt{2}-\frac{1015}{288}\,\mu^3+\frac{803}{324}\,\mu^3\sqrt{2}$
$\frac{\pi}{2}$	$1-7/3\,\mu+\frac{49}{18}\,\mu^2-\frac{343}{324}\,\mu^3$
π	$1-16/3\,\mu+\frac{128}{9}\,\mu^2-\frac{1024}{81}\,\mu^3$

Table 12
Values of $g(\theta,\mu)$ for some θ, SSPRK(3,2) (explicit scheme A from (33)), fourth order space discretization (9).

The point where the stability region intersects the negative real half-line is located at $x \approx -4.519$.

The dissipativity analysis for A yields the amplification factors for the standard three-point space discretization (8) and the fourth order stencil (9). These amplification factors are evaluated at the points $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \pi\}$ in Tables 11 and 12, respectively.

For the three-point space discretization (8), the first positive zero of $g(\pi,\mu)$ is ≈ 0.8968, where the function changes its sign, and $g(\pi,\mu) = -1$ for $\mu \approx 1.129$. The first positive zero of $g(\pi,\mu)$ is ≈ 0.6726 for the fourth-order space discretization (9), where the function changes its sign, and $g(\pi,\mu) = -1$ at $\mu \approx 0.8474$.

References

[1] U. Ascher, S. Ruuth, R. Spiteri, Implicit–explicit Runge–Kutta methods for time-dependent partial differential equations, Appl. Numer. Math. 25 (1997) 151–167.

[2] U. Ascher, S. Ruuth, T. Wetton, Implicit–explicit methods for time-dependent partial differential equations, SIAM J. Numer. Anal. 32 (3) (1995) 797–823.

[3] J. Butcher, The Numerical Analysis of Ordinary Differential Equations., John Wiley, Chichester, U.K., 1987.

[4] V. Canuto, Turbulence in astrophysical and geophysical flows, in: W. Hillebrandt, F. Kupka (eds.), Interdisciplinary Aspects of Turbulence, vol. 756 of Lecture Notes in Physics, Springer Verlag, 2009, pp. 107–160.

[5] V. Canuto, Stellar mixing. II. Double diffusion processes, Astron. Astrophys. 528 (2011) A77.

[6] S. Chandrasekhar, Hydrodynamic and Hydromagnetic Stability, Clarendon Press, Oxford, 1961.

[7] T. Cowling, The non-radial oscillations of polytropic stars, Mon. Not. Roy. Astron. Soc. 101 (1941) 367–375.

[8] R. Donat, A. Marquina, Capturing shock reflections: An improved flux formula, J. Comput. Phys. 125 (1996) 42–58.

[9] E. Fehlberg, Klassische Runge-Kutta-Formeln vierter und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme, Computing 6 (1970) 61–71.

[10] L. Ferracina, M. Spijker, Strong stability of singly-diagonally-implicit Runge–Kutta methods, Appl. Numer. Math. 58 (2008) 1675–1686.

[11] J. Ferziger, M. Perić, Computational Methods for Fluid Dynamics, Springer Verlag, Berlin–Heidelberg–New York, 2002.

[12] S. Gottlieb, L.-A. Gottlieb, Strong stability preserving properties of Runge–Kutta time discretization methods for linear constant coefficient operators, J. Sci. Comput. 18 (2003) 83–109.

[13] S. Gottlieb, D. Ketcheson, C.-W. Shu, High order strong stability preserving time discretizations, J. Sci. Comput. 38 (2009) 251–289.

[14] H. Grimm-Strele, Numerical solution of the generalised Poisson equation on parallel computers, Master's thesis, University of Vienna, available from `http://othes.univie.ac.at/9200/` (March 2010).

[15] E. Hairer, S. Nørsett, G. Wanner, Solving Ordinary Differential Equations I, Springer-Verlag, Berlin–Heidelberg–New York, 1987.

[16] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II, Springer-Verlag, Berlin–Heidelberg–New York, 1991.

[17] N. Happenhofer, Simulation of low mach number fluids, Master's thesis, University of Vienna, available from `http://othes.univie.ac.at/10551/` (June 2010).

[18] K. Heun, Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen., Zeitschrift f. Mathematik u. Physik 45 (1900) 23–38.

[19] I. Higueras, Radius and regions for arbitrary γ, private communication.

[20] I. Higueras, Representations of Runge–Kutta methods and strong stability preserving methods, SIAM J. Numer. Anal. 43 (2005) 924–948.

[21] I. Higueras, Strong stability for additive Runge–Kutta methods, SIAM J. Numer. Anal. 44 (2006) 1735–1758.

[22] I. Higueras, Characterizing strong stability preserving additive Runge–Kutta methods, J. Sci. Comput. 39 (2009) 115–128.

[23] A. Hujeirat, R. Rannacher, On the efficiency and robustness of implicit methods in computational astrophysics, New Astronomy Reviews 45 (2001) 425–447.

[24] H. Huppert, P. Linden, On heating a stable salinity gradient from below, J. Fluid Mech. 95 (1979) 431–464.

[25] C. Kennedy, M. Carpenter, Additive Runge–Kutta schemes for convection–diffusion–reaction equations, Appl. Numer. Math. 44 (2003) 139–181.

[26] R. Kippenhahn, A. Weigert, Stellar Structure and Evolution, 3rd print. of 1st ed., Springer-Verlag, 1994.

[27] O. Koch, F. Kupka, B. Löw-Baselli, A. Mayrhofer, F. Zaussinger, SDIRK methods for the ANTARES code, ASC Report 32/2010, Inst. for Anal. and Sci. Comput., Vienna Univ. of Technology, available at `http://www.asc.tuwien.ac.at/preprint/2010/asc32x2010.pdf` (2010).

[28] J. Kraaijevanger, Contractivity of Runge–Kutta methods, BIT 31 (1991) 482–528.

[29] F. Kupka, J. Ballot, H. Muthsam, Effects of resolution and Helium abundance in A star surface convection simulations, Comm. in Asteroseismology 160 (2009) 30–63.

[30] N. Kwatra, J. Su, J. Grétarsson, R. Fedkiw, A method for avoiding the acoustic time step restriction in compressible flow, J. Comput. Phys. 228 (2009) 4146–4161.

[31] X. Liu, S. Osher, Convex ENO high order multi-dimensional schemes without field by field decomposition or staggered grids, J. Comput. Phys. 142 (1998) 304–330.

[32] X.-D. L. M. Kang, R.P. Fedkiw, A boundary condition capturing method for multiphase incompressible flow, J. Sci. Comput. 15 (2000) 323–360.

[33] H. Muthsam, W. Göb, F. Kupka, W. Liebich, Interacting convection zones, New Astronomy 4 (1999) 405–417.

[34] H. Muthsam, W. Göb, F. Kupka, W. Liebich, J. Zöchling, A numerical study of compressible convection, Astron. Astrophys. 293 (1995) 127–141.

[35] H. Muthsam, F. Kupka, B. Löw-Baselli, C. Obertscheider, M. Langer, P. Lenz, ANTARES — A Numerical Tool for Astrophysical RESearch with applications to solar granulation, New Astronomy 15 (2010) 460–475.

[36] L. Pareschi, G. Russo, Implicit–explicit Runge–Kutta schemes and application to hyperbolic systems with relaxation, J. Sci. Comput. 25 (2005) 129–155.

[37] S. Ruuth, R. Spiteri, High-order strong-stability-preserving runge-kutta methods with downwind-biased spatial discretizations, SIAM J. Numer. Anal. 42 (2004) 974–996.

[38] M. Schwarzschild, M. Härm, Evolution of very massive stars, Astrophys. Jour. 128 (1958) 348–360.

[39] C.-W. Shu, Total-variation-diminishing time discretizations, SIAM J. Sci. Statist. Comput. 9 (1988) 1073–1084.

[40] C.-W. Shu, Essentially non-oscillarory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws, Tech. Rep. ICASE 97-65, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA (1997).

[41] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock–capturing schemes, J. Comput. Phys. 77 (1988) 439–471.

[42] R. Spiteri, S. Ruuth, A new class of optimal high-order strong-stability-preserving time discretization methods, SIAM J. Numer. Anal. 40 (2002) 469–491.

[43] H. Spruit, The rate of mixing in semiconvective zones, Astron. Astrophys. 253 (1992) 131–138.

[44] H. Spruit, Å. Nordlund, A. Title, Solar convection., Ann. Rev. Astron. Astrophys. 28 (1990) 263–301.

[45] J. Strikwerda, Finite Difference Schemes and Partial Differential Equations, 2nd ed., SIAM, Philadelphia, PA, 2004.

[46] J. Turner, Multicomponent convection, Ann. Rev. Fluid Mech. 17 (1985) 11–44.

[47] A. Weiss, W. Hillebrandt, H. Thomas, H. Ritter, Cox and Giuli's Principles of Stellar Structure, Advances in Astronomy and Astrophysics, 2nd ed., Cambridge Scientific Publishers Ltd., Cambridge, U.K., 2004.

[48] F. Zaussinger, Numerical simulation of double–diffusive convection, Ph.D. thesis, University of Vienna, available from `http://othes.univie.ac.at/13172/` (December 2010).

[49] F. Zaussinger, H. Spruit, Semiconvection, submitted to Astron. Astrophys. (for a preprint see arXiv:1012.5851v2) (2011).